Traffic-Sign Detection and Classification in the Wild

0. Abstract

Mặc dù đã đạt được một số kết quả đầy hứa hẹn trong việc detect và phân loại biển báo giao thông, một số ít các công trình đã đưa ra giải pháp đồng thời cho cả tác vụ này áp dụng cho các hình ảnh ngoài đời thực. Đây là một trong số các công trình như vậy.

Đầu tiên, chúng tôi xây dựng một benchmark hiệu quả với tập khá lớn với 100,000 các tấm ảnh góc rộng từ Tencent Street View, khác xa các benchmark trước đó cho vấn đề này. Từ đó ta có 100,000 tấm ảnh chứa khoảng 30,000 biển báo giao thông, trong những điều kiện ánh sáng và thời tiết khác nhau. Mỗi biển báo trong tập được gán nhãn, được đánh dấu khung và mặt nạ pixel.

Chúng tôi đã trình bày một mạng CNN mạnh mẽ có thể detect được biển báo giao thông một cách đồng thời. Hầu hết các giải pháp trước đây đều tập trung nhận diện vật thể lớn, mà bỏ qua hoặc nhận diện không tốt các vật thể nhỏ. Kết quả thực nghiệm cho thấy mạng CNN được đề xuất sẽ là một sự thay thế ưu việt.

1. Introduction

Hiểu được Scene (ngữ cảnh) là mục tiêu tối thượng của Thị giác máy tính, trong đó có xác định và phân loại đối tượng lớn nhỏ khác nhau xuất hiện trong ảnh là một trong những nhiệm vụ thiết yếu. Gần đây thì DeepLearning đã mang đến những phương thức mới đột phá trong việc xử lý hình ảnh và tiếng nói. Trong đó một biến thể của DL là Mạng CNN cho thấy sức mạnh của nó trong bài toán phân loại hình ảnh, localization và detection. Có 2 benchmark phổ biến để đánh giá độ hiệu quả trong detection là PASCAL VOC[7] và ImageNet ILSVRC [20]. Trong các tập data này, đối tượng cần detect thường chiếm kích thước lớn (khoảng hơn 20%) trong ảnh. Tuy nhiên, thực tế đối tượng có thể là một phần nhỏ trong ảnh, chẳng hạn như là biển báo giao thông, với size thông thường khoảng 80x80p, và thường chiếm chỉ 0.2% bức ảnh. Trong thực tế, nhu cầu nhận dạng và phân loại các thực thể kích thước nhỏ là rất cần thiết, vì vậy cần phải có giải pháp và phương pháp đánh giá các thuật toán có nhận diện tốt hay không khi mà đối tượng chỉ là một phần phụ nhỏ trong ảnh.

Ta có thể chia biển báo giao thông ra nhiều loại tùy vào chức năng, và mỗi loại lại được chia thành nhiều loại con nhỏ hơn, có cùng kiểu màu sắc nhưng chi tiết lại khác nhau. Giải pháp này có thể được phân nhỏ thành 2 pha: nhận dạng sau đó phân lớp. Bước nhận dạng sử dụng những thông tin được chia sẻ, đã được học để dự đoán khung nào có chứa biển báo giao thông, sau đó bước phân loại sẽ gán nhãn cho khung đó nếu có thể.

Kể từ khi có tập data về biển báo đầu tiên của Đức [24,25] đến nay, có nhiều nhóm nghiên cứu đã làm về 2 pha này (detection & classification), và đạt đến gần 100% độ chính xác (recall & precision for detection) và 99.67% (classification). Tuy nhiên các tập này lại không đại diện cho hầu hết các trường hợp thực tiễn. Trong GTFDB [25], thuật toán chỉ nhận dạng biển báo của 1

trong 4 loại chính. Trong GTSRB [24], biển báo lại xuất hiện trong hầu hết các bức ảnh, và thuật toán chỉ việc phân loại biển báo trong ảnh mà thôi. Hơn nữa, lại không có các mẫu negative (các đối tượng nhiễu) gây gián đoạn việc phân loại. Trong thực tế biển báo chỉ chiếm rất nhỏ trong ảnh, thường nhỏ hơn 1%. Bởi vậy các vùng ứng viên bị nhỏ hơn tiêu chuẩn cần thiết trong các benchmark PASCAL VOC hay ImageNet ILSVRC. Ngoài ra, thuật toán phải lọc ra được các trường hợp negative tiềm năng trong lúc giữ lại các trường hợp nhận dạng đúng. Chúng tôi đã xây dựng một benchmark mới thực tế hơn và còn dùng nó để đánh giá một hướng tiếp cận bằng CNN, cụ thể:

- Tập dùng để benchmark đồ sộ và thực tế hơn các tập sẵn có, với độ phân giải gấp nhiều lần và nhiều môi trường, độ sáng và ảnh hưởng của thời tiết đa dạng hơn. Benchmark này được chú thích bằng một mặt nạ pixel cho mỗi biển báo, cũng như xác định khung đối tượng trong ảnh. Ta gọi tập này là Tsinghua-Tencent 100K.
- Đã huấn luyện được 2 mạng CNN cho việc nhận dạng biển báo, và đồng thời nhận dạng và phân lớp biển báo đó luôn. Các phép đánh giá trên benchmark này cho thấy hiệu quả mạnh mẽ khi kết hợp hai mạng này với nhau.

2. Related work - đôi nét về các nghiên cứu liên quan đến chủ đề này

a. Traffic sign Classification

Trước CNN, có nhiều phương pháp nhận dạng đối tượng được áp dụng cho phân loại biển báo, ví dụ như SVMs và "sparse representations". Gần đây thì CNN đã cho thấy hiệu quả tốt trong các bài toán phân loại đơn giản khi test trên benchmark GTSRB. Các CNN này có thể là "committe of CNN", multi-scale CNNs và CNN-with-hinge loss function, đều đã được độ chính xác 99.65%, tốt hơn cả con người. Tuy nhiên như đã nói ở trên, vẫn chưa đủ để ứng dụng thực tiễn.

b. Object detection by CNNs

CNNs cũng được dùng cho phân loại hình ảnh nhờ vào khả năng nhận dạng nhanh chóng đối tượng. Trong "OverFeat", Sermanet đã chứng minh CNN vốn dĩ rất hiệu quả khi được dùng trong một khung trượt, và đã minh họa một mạng neural có khả năng xác định khung đối tượng và gán nhãn nó luôn.

Một chiến lược khác là dùng CNNs để tìm ra generic object proposals trước tiên, sau đó mới phân lớp các ứng viên ứng với proposal đó. R-CNN là mạng đầu tiên dùng cách này, nhưng chạy rất chậm. Nguyên nhân là do việc tạo ra các proposal cho đối tượng quá lâu. Thuật toán Selective search cũng mất 3 giây để gen ra 1000 proposal cho tập hình Pascal VOC 2007, và EdgeBoxes xịn xò hơn cũng tốn 0.3s. Nguyên nhân thứ hai là do sử dụng CNN cho từng proposal ứng viên một, nên gây chậm cho toàn quá trình.

Để cải thiện điều này, mạng gộp kim tự tháp SPP-Net đã dùng một "convolutional-feature-map" chung cho tất cả các hình và trích xuất các vector đặc trưng từ các shared-feature-map cho từng proposal. Việc này giúp tăng tốc xử lý R-CNN ở trên lên 100 lần.

Sau đó mạng Fast R-CNN được đề xuất, sử dụng một softmax-layer đặt ngay trên đầu mạng thay cho SVM ban đầu. Nếu bỏ qua thời gian proposal, thì cách này giúp xử lý mỗi hình tốn 0.3s. Và trong Faster R-CNN, Ren đề xuất mạng RPN dùng convolutional-feature map để gen các proposal. Điều này cho phép quá trình generate có thể share các đặc trưng chung của tấm hình gốc (full-image) cho các hình nhỏ hơn, giúp làm tăng tốc độ generate hơn.

Thay cho việc tìm ra các proposal một cách thủ công, Szegedy đã cải tiến dựa trên một phương pháp generate dựa vào data đã có, cũng như cải tiến kiến trúc mạng để tối ưu đến được 50fps lúc test, với độ chính xác không kém cạnh.

Tuy nhiên, tất cả các mạng trình bày trên đây đều sử dụng phép đánh giá trên PASCAL VOC và ILSVRC, dùng cho đối tượng lớn trong bức ảnh.

3. Benchmark

Phần này gồm có 3 ý chính : data được lấy ở đâu, làm sao annotate data, và rốt cuộc data chứa những thông tin gì.

a. Data collection

Trong khi các tập data phổ thông như ImageNet và COCO được lấy từ các nguồn ảnh từ Internet bằng các search engine, mà ở đó khá ít user tải ảnh từ thực tế cuộc sống mà có biển báo giao thông như khi ta đi trên đường, hoặc nếu có thì cũng chỉ là vô tình và ngẫu nhiên, mà không được đánh tag đúng với loại biển báo chứa trong ảnh đó. Vì vậy phải có cách thu thập data khác cách này. Ngoài ra để bắt chước một ngữ cảnh giống thực tế, các tấm ảnh không chứa biển báo cũng được thêm vào, để đánh giá xem hệ thống có phân biệt được biển báo thật và các đối tượng na ná giống, gần giống hay không. Vậy cho nên chúng tôi quyết định thu thập dữ liệu từ Tencent Street Views.

Hiện tại Tencent Street Views phủ khoảng 300 thành phố ở Trung Quốc và các tuyến đường liên thông giữa các thành phố. Các góc quay chụp góc rộng được chụp với camera 6 SLR và sau đó ghép nối lại với nhau. Các kỹ thuật xử lý ảnh như điều chỉnh độ phơi sáng cũng được dùng. Các ảnh này được chụp từ phương tiện giao thông và các thiết bị đeo với chu kỳ khoảng 10 phút. Nhờ sự tự nhiên khi chụp này, giúp mang lại 2 lợi ích:

- Thứ nhất, các ảnh được chụp liên tiếp thì liên kết đồng nhất với nhau. Không như trong GTSRB, khi các biển báo được trích xuất từ một phân đoạn video, tức hình biển báo sẽ rất giống nhau ở mọi hình.
- Thứ hai, có được các tấm ảnh liên tục về một biển báo sẽ tăng độ chính xác cho việc xác định loại biển báo đó, ví dụ khi biển báo bị che khuất hoặc bị mờ một phần nào đó, có thể được nhận dạng từ những tấm hình trước đó hoặc sau đó trong chuỗi hình liên kết này.

Để chuẩn hóa các tấm ảnh, 25% phía trên và 25% phía dưới ảnh sẽ được bỏ đi vì không có dính biển báo. Phần giữ lại sẽ được cắt theo chiều dọc thành 4 tấm ảnh con (Hình 1).

Chúng tôi chọn ra 10 vùng từ 5 thành phố khác nhau ở TQ (gồm cả nội thành lẫn ngoại ô) và tải 100 nghìn tấm ảnh góc rộng như vậy từ Tencent Data Center.

b. Data annotation

Các tấm ảnh được chọn sau đó được annotate thủ công. Ở TQ thì biển báo tuân theo chuẩn quốc tế, và có thể phân lớp thành 3 hoặc 4 loại:

- biển cảnh báo (tam giác vàng viền đen)
- biển cấm (tròn trắng viền đỏ)
- biển hiệu lệnh (tròn xanh hình trắng)
- các loại biển báo khác ít phổ biến

Trong quá trình gán nhãn, chúng tôi ghi dấu lại khung bao bounding-box, các đỉnh của khung và gán nhãn cho biển báo đó. Để xác định mặt nạ pixel cho biển báo, chúng tôi dùng 2 mode: đa giác và elip. Trong mode đa giác, chúng tôi đánh dấu các đỉnh của đối tượng ứng với đỉnh của đa giác. Trong mode elip, chúng tôi đánh dấu tùy ý các đỉnh này dọc theo đường elip. Đối với biển báo tam giác, chúng tôi chỉ đánh dấu 3 đỉnh/cạnh. Đối với các biển báo bị méo mó, chúng tôi đánh dấu thêm các phân đoạn bổ sung. Đối với các biển hình tròn, nó sẽ là elip khi không bị che đi, nên chúng tôi đánh dấu 5 đỉnh trước để hậu xử lý sau. Các trường hợp phức tạp thường gặp là ở ví dụ hình 5. Khi bị che đi như thế này, chúng tôi đánh dấu khung bao cả bên ngoài, khung bao đa giác bao và cả khung bao elip nếu có, sau đó giao chúng lại để lấy khung bao cuối cùng.

c. Dataset statistics

Tiếp đến chúng tôi có được 100 nghìn các ảnh được crop và lược bỏ các tấm chỉ toàn background. Cuối cùng cho ra được 10 nghìn tấm có tổng cộng 30 nghìn biển báo. Mặc dù các ảnh đều toàn lấy ở Trung Quốc, vẫn có tồn tại một số lớp khác nhau trong tập ảnh vì kích thước và độ phổ biến khác nhau. Điều này là không thể tránh khỏi, ví dụ như biển cảnh báo thận trọng khi đi đường núi thì ít phổ biến, xuất hiện ít hơn. Nói cách khác, tần suất xuất hiện của các loại biển báo là không giống nhau. Hình 6 và 7 thể hiện điều này.

Sau cùng, bộ data chúng tôi xây dựng được cung cấp đánh dấu rất chi tiết cho từng biển báo: khung bao, mặt nạ pixel và phân lớp của nó. Các biển báo được phân bố ra nhiều lớp, và số lượng mỗi lớp là khác nhau. Hình sử dụng trong tập này có độ phân giải là 2048x2048, kèm các biến thể của ảnh theo độ sáng và điều kiện thời tiết. Đây là các thông tin, dữ liệu cơ bản cho cả 2 pha nhận dạng và phân lớp các đối tượng biển báo nhỏ sẽ được trình bày sau đây.

4. Neural network

Chúng tôi huấn luyện tổng cộng 2 model, một model cho việc nhận dạng riêng lẻ, và một model cho nhận dạng và phân lớp một cách đồng thời. Hai model này có cấu trúc khá tương đồng nhau ngoại trừ một nhánh ở layer cuối cùng

a. Architecture

Trong [12], Huval đánh giá hiệu quả của mạng CNNs dựa trên việc nhận dạng đường và phương tiện. Họ dùng framework OverFeat với một bước hồi quy để tìm ra khung bao đối tượng. Mạng này hoàn toàn là convolutional, và layer cuối cùng được phân thành 2 luồng:

- pixel layer. Mỗi kết quả đầu ra của pixel layer thể hiện xác suất của vùng pixel 4x4 trong ở ảnh input có chứa đối tượng.

- bounding-box layer: mỗi kết quả output thể hiện khoảng cách giữa vùng được chọn này so với 4 phía của bounding-box của đối tượng.

Phép đánh giá được thực hiện trên 1.5 giờ video quay trên cao tốc. Mặc dù mạng này nhận dạng cực tốt các phương tiện, tuy nhiên mạng không thể thích ứng ngay với việc nhận dạng nhiều lớp đối với đối tượng nhỏ như yêu cầu đề bài. Tuy nhiên, chúng tôi đã xây dựng lại dựa trên mạng này cùng với một số điều chỉnh đáng chú ý:

- Đầu tiên, chúng tôi làm cho mạng được phân nhánh sau layer thứ 6, khác với bản gốc được phân nhánh sau layer thứ 7. Trong quá trình thí nghiệm, chúng tôi phát hiện ra sự điều chỉnh này làm cho mạng nhanh hơn. Như đã nêu trong [23, 26], mạng càng nhiều lớp thì hiệu quả càng cao. Nếu ta phân nhánh càng sớm, thì có cơ hội tăng độ hiệu quả, tuy nhiên làm tăng chi phí huấn luyện và bộ nhớ GPU. Cách này có vẻ không tối ưu chi phí.
- Một điều chỉnh nữa là mạng của chúng tôi phân làm 3 luồng thay vì 2 như bản gốc. Ngoài 2 luồng chính, có thêm một layer để gán nhãn, có output là một vector n chiều, ứng với xác suất đối tượng thuộc n lớp cụ thể. Điều này làm cho mạng thực hiện đồng thời việc nhận dạng và phân lớp biển báo luôn.

Kiến trúc mạng đề xuất này được minh họa trong hình 8. Các chi tiết cụ thể được mô tả trong source code. Mô hình được cài đặt trên framework Caffe. Sau khi bỏ đi layer gán nhãn ở layer thứ 8, mạng có thể dùng để chỉ nhận dạng biển báo.

b. Training

Do số lượng mẫu không đồng đều giữa các lớp với nhau trong các loại biển báo, chúng tôi đã sử dụng kỹ thuật gia tăng dữ liệu (data augmentation) trong quá trình huấn luyện. Chúng tôi đơn giản là bỏ qua các lớp có ít hơn 100 mẫu, còn lại sau cùng là 45 lớp. Nếu số mẫu từ 100 đến 1000 mẫu thì gia tăng cho lên 1000 mẫu. Còn lại không đổi.

Để gia tăng được số mẫu như trên, chúng tôi dùng template chuẩn cho mỗi phân lớp biển báo, rồi xoay tròn ngẫu nhiên một khoảng [-20 độ,20 độ], scale ngẫu nhiên trong khoảng [20%,200%], và thêm ngẫu nhiên một ít độ méo mó sao cho tự nhiên nhất.

Sau đó chúng tôi chọn thủ công các hình không có biển báo và trộn lẫn vào trong tập template đã gia tăng, và thêm một ít nhiễu.

5. Result

Trong thí nghiệm này, cả hai pha huấn luyện và kiểm thử đều thực hiện trên Linux PC với chip Intel Xeon E5-1620 CPU,2 NDIVIA tesla K20 GPU và 32GB RAM. Khoảng 10 nghìn bức ảnh góc rộng chứa biển báo, chúng tôi chia ra thành tập huấn luyện và tập kiểm tra với tỉ lệ 2:1 để cung cấp thật nhiều mẫu huấn luyện. 90 nghìn ảnh còn lại được dùng trong pha kiểm tra.

Chúng tôi dùng số liệu đánh giá sử dụng Microsoft COCO benchmark, và chia các biển báo ra 3 loại theo kích thước:

- nhỏ (dưới 32x32)

- vừa (từ 32x32 đến 96x96)
- lớn (hơn 96x96)

Phép đánh giá này có thể cho thấy khả năng nhận dạng khi đối tượng có kích thước khác nhau.

a. Detection

Đến đây ta cân nhắc xem phương thức này hoạt động tốt và nhận dạng tốt đến đâu, bằng cách sử dụng benchmark này. Một đặc trưng phổ biến của các mạng nhận dạng đối tượng hiện nay là nó dựa vào bản proposal tổng quát của đối tượng. Nếu từ đầu đối tượng không thuộc tập proposal, thì các bước sau là vô ích. Kích thước của đối tượng dùng trong phép đo của chúng tôi thì nhỏ hơn nhiều, và phương pháp trích xuất các proposal điển hình thì không hoạt động tốt với các đối tượng nhỏ. Kết quả đo được thể hiện như trong hình. Selective Search, Edge Boxes và MultiScale Combinatorial Grouping (MCG) được đề xuất trong [11] là các phương pháp hiệu quả nhất có thể. Do MCG là memory intensive còn các hình ảnh của chúng tôi có độ phân giải cao, chúng tôi thay bằng BING. Lưu ý rằng Selective Search và Edge Boxes không cần dữ liệu huấn luyện, nên ta có thể đánh giá độ hiệu quả một cách trực tiếp trên tập kiểm tra luôn. Chúng tôi huấn luyện BING với cùng một tập gia tăng như trên.

Kết quả được thể hiện trong hình 9. Lượng recall trung bình của toàn bộ 10 nghìn proposal của 3 phương pháp đều dưới 0.7. Điều này chỉ ra rằng nhận dạng proposal cho tập hình nhỏ không được hiệu quả, ngay cả trong tập ứng viên đủ lớn.

Thay vào đó, ghom hết các biển báo vào thành một loại, chúng tôi đã dùng mạng được đề xuất để huấn luyện và đạt được 84% độ chính xác, và 94% recall với hệ số tương tự Jaccard là 0.5, mà không cần điều chỉnh bộ tham số của nó. Kết quả này vượt trội hơn đáng kể so với các phương pháp đã liệt kê. Hơn nữa, hiệu quả không bị giảm sút khi thực hiện trên nhiều kích thước khác nhau.

Chúng tôi còn chạy kiểm tra trên tập 90 nghìn ảnh không chứa biển báo, và mạng đã xác định cực kỳ chính xác rằng chỉ có nền. Thông tin chi tiết được kèm theo trong tài liệu bổ sung.

b. Simultaneous detection and classification: Phát hiện và phân loại đồng thời

Đến đây chúng tôi chuyển sang bài toán kết hợp giải quyết cả 2 nhiệm vụ là nhận dạng biển báo đồng thời phân loại chúng. Chúng tôi so sánh kết quả của mạng này với Fast R-CNN. Chúng tôi cũng generate ra 10 nghìn proposal cho mỗi hình khi chúng tôi chạy test trên Fast R-CNN. Kết quả được mô tả trong hình 10. Kết quả cho thấy sự vượt trội, đặc biệt đối với các hình nhỏ. Các thông số so sánh được cho ở bảng 2 kèm theo.

6. Conclusion

Kết luận, chúng tôi đã dựng được một benchmark cho việc nhận dạng và phân loại biển báo đồng thời nhau. So sánh với các phương pháp khác, thì tập hình ảnh có nhiều biến thể hơn, và biển báo thì nhỏ hơn nhiều. Tập ảnh cũng nhiều hơn và độ phân

giải cũng cao hơn. Xa hơn nữa, phân đoạn pixel-wise của biển báo sẽ được cung cấp. Phương pháp này cung cấp một thách thức mới cho cộng đồng nhận dạng biển báo. Chúng tôi đã huấn luyện 2 mạng khác nhau: một gom hết các biển báo thành một lớp, coi như là nhận diện biển báo. Mạng còn lại thì vừa nhận diện vừa phân lớp biển báo. Cả hai mạng này đều hiệu quả vượt trội so với các phương pháp trước đó, có thể được dùng cho các nghiên cứu trong tương lai.