Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00022

EMPIRICAL EVALUATION OF STATE-OF-THE-ART OBJECT DETECTION METHODS FOR DOCUMENT IMAGE UNDERSTANDING

Nguyen D. Vo[1], Khanh Nguyen[2], Tam V. Nguyen[3] , Khang Nguyen[2]

[1] University of Science, VNU-HCM

[2] University of Information Technology, VNU-HCM

[3]University of Dayton, USA

voduynguyen2901@gmail.com, khanhnd@uit.edu.vn, tamnguyen@udayton.edu, khangnttm@uit.edu.vn

ABSTRACT: *The majority of online documents such as research papers, articles, and magazines is publicly available in the image form due to the copyright issue. Document image understanding is the task of deriving a high level presentation of the contents of a document image, which involves several phases, mainly including page segmentation (or block segmentation), blocks classification (or blocks labeling) and several operations for processing text, tables, graphics, figures, formulas, etc. Our objective focuses on the first two phases of document image understanding, namely, locating the logical objects in document pages. This process is valuable for a variety of document image analysis applications. To this end, we evaluate different state-of-the-art object detection methods based on computer vision for the task. Through our extensive experiments, we report findings/comments from the off-the-shelf object detectors and streamline several potential directions for the future work.*

Keywords*: Page Object Detection, Document Image Understanding.*

I. INTRODUCTION

Document Image Understanding (DIU) is an interesting research area with a large variety of challenging problems, which has been receiving increasing attention not only from the document analysis and recognition community, but also from the database and information extraction (IE) communities. Page Object Detection (POD) is to detect the specific page objects (e.g. tables, formulas, figures (including charts)) in document images. Objects have a variety of shapes, sizes, and content. Diversity in the location of components inside the object (formula with multiple symbols, one line or multiple lines). ICDAR 2017 POD Competition[1] release a dataset on paper object detection. The competition focuses on the first two phases of document image understanding, locating the logical objects in document pages. The targeted page objects of this competition includes formulas, tables, and images or graphs (including charts). Even though page object detection is a novel and challenging problem, we still can still transfer the contemporary works in object detection [1, 2, 3] into the new problem. Therefore, in this paper, our objective is to evaluate state-of-the-art methods for detecting objects in images. We evaluate deep learning methods. The extensive experiments show some insights from the state-of-the-art methods on the challenging problem.

The rest of the paper is organized as follows. In Section 2, we give a detail presentation of the state-of-the-art object detection includes in our comparison. Section 3 then presents the evaluation and discusses the results obtained for different detection methods. Section 4 concludes the paper.

II. RELATED WORKS

In object detection using feature extraction, there are two main approaches using handcrafted features and deep learned features as follows.

A. Handcrafted Feature-based Works

Histogram of gradients (HOG), which is first proposed by Dalal and Triggs [6], is one of the most successful hand-crafted features for object detection and recognition. The idea behind is to calculate the sum of oriented gradient vectors over local regions. This gradient-based representation method makes HOG very effective under various illumination changes and small deformations. DPM is inherited from HOG feature, and then it extends HOG for representing objects with considerable variations in shape and appearance.

Later, Deformable Part Model (DPM) was proposed by Felzenszwalb and et al. [1]. It detects and locates objects in images based on local parts of objects. DPM consists of two main submodels: (1) a model for the visual appearance of each part and (2) a geometric model that captures spatial relationships between the parts. Maximum likelihood estimation products the parameters of DPM. The success of DPM model comes from both the robustness of HOG feature and the flexibility of geometric part-based representation. The number of parts for object categoríes can also be specified for better representing their appearance.

[1] http://www.icst.pku.edu.cn/cpdp/ICDAR2017_PODCompetition/index.html

B. Deep Learning Methods

R-CNN

Regarding deep learning methods, the Region-based Convolutional Network method (R-CNN) was first proposed by Girshick et al. [7]. It achieves excellent object detection accuracy by using a deep ConvNet to classify object proposals. R-CNN is just the following steps: 1) Proposing multiple regions in an image (~ 2 thousand regions), 2) Classifying each region and 3) Filter the results using non-max suppression. R-CNN creates region proposals, using a process called Selective Search [8]. The classification of each region is done by first extracting features using a CNN. Since the CNN has a fixed-sized input (and the size of the regions vary), R-CNN warps the region to a standard square size and passes it through to a modified version of AlexNet. R-CNN classifíes with a Support Vector Machine (SVM) trained for each class. The last step is a greedy non-maximum suppression. R-CNN runs a simple linear regression on the region proposal to generate tighter bounding box coordinates to get our final result. R-CNN has notable drawbacks: Training is a multi-stage pipeline, training is expensive in space and time, object detection is slow. R-CNN is slow because it performs a ConvNet forward pass for each object proposal, without sharing computation.

Fast R-CNN

Fast R-CNN was proposed by Girshick [2]. It is based on a proposal detection net for object detection tasks. Fast R-CNN is a single stage training algorithm that classifies object proposals and refines their localisation. The input of Fast R-CNN is an image and a set of regions of interest (RoI). The network uses several convolutional (conv) and max pooling layers to produce a feature map of the entire image. Normally there are about 2000 RoIs, which are determined by proposal methods like Selective Search. The pooling layer will extract a fixed-length feature vector from the feature map of each region of interest. Each vector feeds into a sequence of fully connected layers (FCs). This produces two output vectors for each RoI: one to estimates the object class and another to locates RoI. The softmax function produces probability over K object classes plus a catch-all "background" class. The position of RoI is a set of four real-valued numbers. This method has several advantages: higher detection quality (mAP) than R-CNN [2]; training is single-stage, using a multi-task loss; training can update all network layers; no disk storage is required for feature caching.

Faster R-CNN

Meanwhile, Faster-RCNN was proposed by Ren et al. [3]. It inherited from Fast-RCNN mainly focus on finding region proposal with CNN, and share features with object classification task followed by region proposals. Faster R-CNN is composed of two modules. The first module is a Region Proposal Network (RPN), and the second module is the Fast R-CNN detector [2]. The input of RPN is an image of any size and the output of RPN is a rectangular object proposal with objectness score. RPN is fully convolutional. RPN is trained end-to-end to generate high-quality region proposals. The Fast R-CNN detector uses the proposed regions for detection. The entire system is a single, unified network for object detection [3]. Figure 1 illustrates the deep networks proposed in Fast RCNN and Faster RCNN.

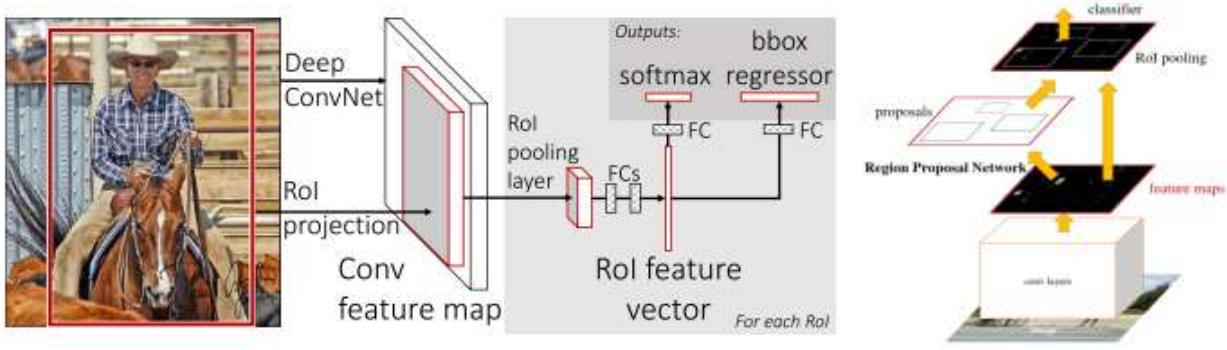

Figure 1. Two state-of-the-art deep learning methods: (left) Fast R-CNN architecture [2], and (right) Faster R-CNN architecture. [3]

In this paper, we will conduct intensive experiments on the state-of-the-art deep learning methods, namely, Fast RCNN and Faster RCNN since they clearly outperformed hand-crafted feature-based methods in literature [2, 3]. In particular, we are evaluating different network structures for deep feature methods, on the ICDAR2017 POD Competition dataset.

III. EVALUATION

A. Dataset

The POD dataset [4] consists of three classes: figure, table, formula. In total, the data consists of 2417 images. Training set includes 1600 images and testing set includes 817 images. We train and evaluate a multi-class classifier and report average accuracy over all classes as a performance measure.

B. Performance Metrics

We used the code is available [2, 3], trained and evaluated on POD dataset. The Intersection over Union (IOU) is an evaluation metric used to measure the accuracy of an object detector on a particular dataset. IoU measure gives the similarity between the predicted region and the ground-truth region. The IOU is calculated as follows:

$$IOU = \frac{S_i \cap S_j}{S_i \cup S_j} \quad (1)$$

where S_i denote the region detected by a participant and S_j denote the corresponding region described in the ground truth file. The IOU is calculated as follows:

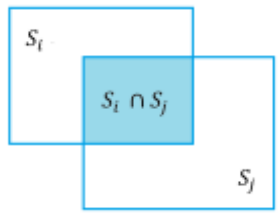

Figure 2. IOU calculation illustration.

As shown in Fig. 2, $S_i \cap S_j$ denotes the area of the intersection of S_i and S_j and $S_i \cup S_j$ denotes the area of the union of S_i and S_j. The IOU is then calculated as the following equation.

$$IOU = \frac{S_i \cap S_j}{S_i + S_j - S_i \cap S_j} \quad (2)$$

The Average Precision (AP) [5] is a common metric to measure for each class. The mean Average Precision (mAP) is computed by taking the average over the APs of all classes. Recall is defined as the proportion of all positive examples ranked above a given rank. Precision is the proportion of all examples above that rank which are from the positive class. The AP summarises the shape of the precision/recall curve, and is defined as the mean precision at a set of eleven equally spaced recall levels [0,0.1,…,1]:

$$AP = \frac{1}{11} \times \sum_{r \in \{0,0.1,\ldots,1\}} Pinterp(P) \quad (3)$$

The precision at each recall level r is interpolated by taking the maximum precision measured for a method for which the corresponding recall exceeds r:

$$Pinterp(P) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r}) \quad (4)$$

where $p(\tilde{r})$ is the measured precision at recall $\tilde{r}$.

C. Performance

For the evaluation, we divide the dataset into 3 splits. For each split, we randomly select 1100 samples as training set whereas the remainder is used as a testing set. We will publish our data splits in order to support the reproduction of future works. Next, we report the results of each method on each dataset split and average performance on all three data splits.

We have to retrain Fast R-CNN with smaller networks (CaffeNet, VGG_CNN_M_1024) and Faster R-CNN with smaller networks (ZF, VGG_CNN_M_1024) on POD Dataset. We still use the parameters that the authors have provided for training phrase [2, 3]. All of models trained on environment: Ubuntu 14.04 64 bits, Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz, 50 GB RAM DDR3, GPU Tesla C2075.

Table 1. Dataset Split 01

Method	Formula (%)	Table (%)	Figure (%)	mAP (%)
Fast R-CNN CaffeNet	2.80	75.30	80.61	52.90
Fast R-CNN VGG_CNN_M_1024	2.47	76.06	82.26	53.60
Faster R-CNN ZF	59.24	95.87	77.87	77.66
Faster R-CNN VGG_CNN_M_1024	62.48	96.49	76.78	78.58

Table 2. Dataset Split 02

Method	Formula (%)	Table (%)	Figure (%)	mAP (%)
Fast R-CNN CaffeNet	0.24	74.66	86.47	54.50
Fast R-CNN VGG_CNN_M_1024	2.16	75.21	87.03	54.80
Faster R-CNN ZF	56.43	97.27	86.06	79.92
Faster R-CNN VGG_CNN_M_1024	61.30	96.05	84.89	80.75

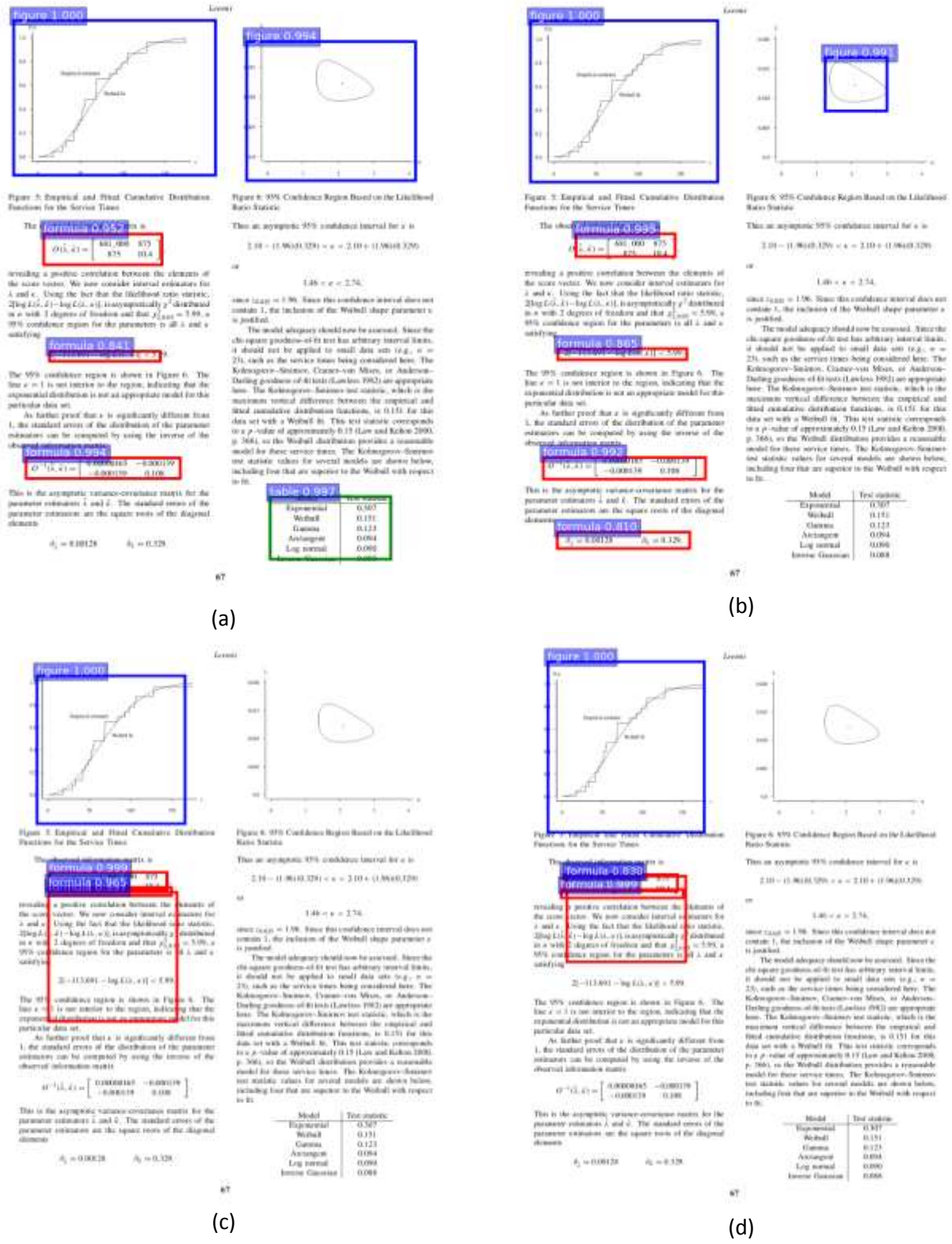

Figure 3. Visualization of object detection results on the POD dataset. Color legend: formulas – red, table – green, figure – blue: (a) Faster R-CNN VGG_CNN_M_1024, (b) Faster R-CNN ZF, (c) Fast R-CNN VGG_CNN_M_1024, (d) Fast R-CNN CaffeNet.

Table 3. Dataset Split 03

Method	Formula (%)	Table (%)	Figure (%)	mAP (%)
Fast RCNN CaffeNet	0.82	69.37	72.06	47.42
Fast RCNN VGG_CNN_M_1024	0.95	70.29	75.10	48.80
Faster RCNN ZF	60.65	94.20	77.75	77.53
Faster RCNN VGG_CNN_M_1024	61.80	94.16	78.01	77.99

Table 4. Average Performance on 3 splits. The best performance of each category is marked as boldfaced.

Method	Formula (%)	Table (%)	Figure (%)	mAP (%)
Fast RCNN CaffeNet	1.29	73.11	79.71	51.37
Fast RCNN VGG_CNN_M_1024	1.86	73.85	**81.46**	52.39
Faster RCNN ZF	58.77	**95.78**	80.56	78.37
Faster RCNN VGG_CNN_M_1024	**61.86**	95.57	79.89	**79.11**

Table 1, 2, 3, and 4 show the results of 4 methods on 3 splits and the average performance, respectively. We observe that the four methods perform consistently on most of 3 splits. It is worth noting that Fast RCNN is sensitive to figures whereas Faster RCNN well detects formulas and tables. In particular, Faster RCNN significantly outperforms Fast RCNN in formula category. The possible reason is that the Faster RCNN possess the region proposal network which is effective to draw the attention to the potential regions. From the results, it is possible to combine both Fast RCNN and Faster RCNN in a unified framework in order to improve the overall performance. Figure 3 illustrates the visual results of different object detection methods.

IV. CONCLUSION

In this paper, we evaluated state-of-the-art deep learning methods for object detection on the POD dataset. Different methods have different pros and cons. Faster R-CNN performs better Fast R-CNN on formula and table class. Fast RCNN is sensitive to figures. Our overall result shows the weakness of state-of-the-art deep learning based detectors on formula objects. This discovery can lead to further researches such as the fusion with OCR frameworks for better detection. In the future, we will consider combining Fast RCNN and Faster RCNN into a unified framework in order to exploit all theỉr advantages.

ACKNOWLEDGEMENTS

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D2-2017-02.

REFERENCES

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI* , 2010.
[2] R. Girshick. Fast R-CNN. In *ICCV*, 2015.
[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
[4] http://www.icst.pku.edu.cn/cpdp/ICDAR2017_PODCompetition/dataset.html
[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results," 2007.
[6] Navneet Dalal, Bill Triggs: Histograms of Oriented Gradients for Human Detection. In *CVPR* 2005, pp. 886-893.
[7] Ross B. Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR* 2014: 580-587
[8] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. In *IJCV*, 2013.

ĐÁNH GIÁ THỰC NGHIỆM CÁC PHƯƠNG PHÁP HIỆN ĐẠI TRONG DÒ TÌM ĐỐI TƯỢNG CHO VIỆC HIỂU ẢNH TÀI LIỆU

Võ Duy Nguyên, Nguyễn Duy Khánh, Nguyễn Văn Tâm, Nguyễn Tấn, Trần Minh Khang

TÓM TẮT: *Phần lớn các tài liệu trực tuyến như các bài báo nghiên cứu, bài báo trên tạp chí khoa học được công bố rộng rãi dưới dạng ảnh do vấn đề bản quyền. Nhiệm vụ của việc hiểu hình ảnh tài liệu là tìm hiểu cách trình bày ở mức cao về nội dung của một hình ảnh tài liệu, bao gồm nhiều giai đoạn, chủ yếu bao gồm phân đoạn trang (hoặc phân đoạn khối), phân loại khối (hoặc dán nhãn khối) và một số thao tác để xử lý văn bản, đồ hoạ, số liệu và công thức. Mục tiêu của chúng tôi trong bài báo này là tập trung vào hai giai đoạn đầu của sự hiểu biết hình ảnh tài liệu, tức là tìm các đối tượng có nghĩa trong các trang tài liệu. Quá trình này có giá trị cho một loạt các ứng dụng phân tích hình ảnh tài liệu. Để thực hiện nhiệm vụ này, chúng tôi tìm hiểu và đánh giá các phương pháp dò tìm đối tượng trong ảnh hiện đại nhất (ở thời điểm viết bài báo này) khác nhau dựa trên thị giác máy tính. Qua các thí nghiệm khác nhau, chúng tôi báo cáo các phát hiện/thảo luận từ các phương pháp dò tìm đối tượng và định hướng một số hướng phát triển tiềm năng cho tương lai.*