Multi-domain Document Layout Understanding Using Few-Shot Object Detection

Pranaydeep Singh, Srikrishna Varadarajan(✉), Ankit Narayan Singh, and Muktabh Mayank Srivastava

ParallelDots, Inc., Lewes, USA
{pranaydeep,srikrishna,ankit,muktabh}@paralleldots.com

Abstract. We try to address the problem of document layout understanding using a simple algorithm which generalizes across multiple domains while training on just few examples per domain. We approach this problem via supervised object detection method and propose a methodology to overcome the requirement of large datasets. We use the concept of transfer learning by pre-training our object detector on a simple artificial (source) dataset and fine-tuning it on a tiny domain specific (target) dataset. We show that this methodology works for multiple domains with training samples as less as 10 documents. We demonstrate the effect of each component of the methodology in the end result and show the superiority of this methodology over simple object detectors. We will open-source the code, trained models, source and target datasets upon acceptance.

Keywords: Object detection · Few-shot · Transfer learning · Domain-invariant · Document layout detection

1 Introduction

The understanding of document layout in terms of finding logical components such as title, paragraphs etc. is a preliminary step towards retrieving information from images of documents. The amount of variability in real-world data coming from multiple domains e.g., documents, invoices etc. makes it a challenging computer vision problem that has intrigued researchers for decades.

Various image processing methodologies [1,7,8] have approached the problem of understanding general documents as well as digitizing historical documents. With the onset of deep learning and data driven approaches, the problem was approached as a pixel-wise segmentation task [12], where each pixel is assigned a class based on its surrounding pixels. In this paper, we explore a new tangent, where the problem is approached as a few-shot object detection problem to identify relevant areas in a document. The motivation is to understand document

P. Singh—Contributed equally.

A. Campilho et al. (Eds.): ICIAR 2020, LNCS 12132, pp. 89–99, 2020.
https://doi.org/10.1007/978-3-030-50516-5_8

structure with as less as 10 tagged examples since digitization tasks generally don't have an abundance of tagged data at hand. However, understanding documents is a complicated task and a dataset consisting of just 10 examples is not enough to train an object detector especially (as they're fully supervised networks requiring large amounts of training data) to understand various structures, like tables or lists.

Hence, we use a transfer learning based approach where we give the network a general understanding of what basic features and structures are contained in a document and then proceed to train on a few-shot task for understanding of specific document types like invoices, resumes, academic papers, journals etc. A few-shot task is described widely as training the model using just a handful of tagged examples.

The initial network which is to be later used for fine-tuning needs to have a wide understanding of document structures and substructures and needs to be trained extensively for it to yield good results when fine-tuned with very less samples. There was no relevant dataset which accommodated these needs and hence, we artificially generated a simple dataset using HTML. We refer to this dataset as Source Dataset. We then proceed to train the described model on this dataset. This trained model now serves as the backbone of all future models we fine-tuned. Using as little as 10, and up to 50 images, we demonstrate that the obtained model learns to understand document structures. We also show that the methodology can be extended to any number of domains with few examples from each. In this paper, we demonstrate the methodology and its application to Invoices and Resume images. We call these domains as Target Domains and the datasets as Target Datasets.

Our contributions consists of the following points

- Applying state of the art object detection techniques for Document Layout Understanding
- Introducing a generalized algorithm which can perform Layout Understanding in multiple domains using just few tagged images (eg: 10).

2 Related Work

There are two sub-parts to the Document Layout Analysis problem

- Geometric Layout Analysis
- Logical Layout Analysis

Geometric Layout Analysis (GLA) is centred around understanding the basic geometric layout of a document, such as skew, page decomposition, text detection etc. Logical Layout Analysis (LLA) focuses on understanding the implied semantic labels in a document, like captions, subheading, table headings etc. GLA has been addressed mainly by image processing methods like Hough Transforms and Binarization. While the GLA problem is as old as Image Processing itself, LLA is a more recent problem and the one which we attempt to solve.

Approaches employed in LLA mainly follow the bottom-up approach. Bottom-up approaches work by finding the smallest entities like words or characters and attempt to aggregate them using a distance metric and an aggregation algorithm like K-Nearest Neighbors or K-D Trees. These approaches [1,7,8] have the advantage of being mostly unsupervised but involve tuning a lot of heuristics. They are also not scalable to document layouts which are different from those the algorithm is tuned on. Comparisons of such approaches are also covered by [6,11]. The most popular and widely used of these approaches is the Docstrum [8] algorithm. While deep learning approaches to LLA also exist, these approaches [3,12] require vast amounts of training data and only learn a fixed set of labels and are thus not useful for few-shot tasks with a wide variety of different labels. We explore an object detection based approach to LLA, which can be fine-tuned on as less as 10 images to understand semantic labels like address, total bill amount, skills, education etc.

Few shot object detection is a task where the tagged training set is very small (say 1–50 images total). Previous work has been explored on the PASCAL VOC/COCO/ImageNet dataset. [2] introduce a Low-shot Object Detector (LSTD) model which is pretrained on a huge Source Dataset and fine-tuned on a small (low-shot) target dataset. The LSTD model is based on Single Shot Detector (SSD) [5] and Faster-RCNN (FRCNN) [10]. Broadly, they use the SSD network to detect foreground segments and a classifier which takes ROIPooled features from the SSD feature maps to classify the detected regions. There are two regularizations introduced by [2], Background Regularization (BGR) and Tk-Regularization (Tk-R) which helps them in learning from just few examples in the target dataset. We use BGR to make the learning of Target domain easier and faster. This is achieved by making the learning of background part in the Target domain easier through this constraint. Tk-R tries to bridge the gap between predictions of the classifier on Source and Target domain. The Source dataset in our case is more basic while [2] assume the Source dataset to be very huge and comprehensive.

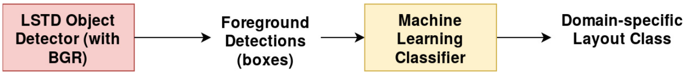

LSTD Object Detector (with BGR)
Foreground Detections (boxes)
Machine Learning Classifier
Domain-specific Layout Class

Fig. 1. Overview of the proposed method

3 Architecture

Our architecture is a two-step object detector. The first step is the detector (inspired from LSTD) which detects the foreground regions and the second step is the ML classifier which predicts the domain-specific layout class.

For the first step, we leverage a better feature extractor for the object detector. We use the Feature Pyramid Networks [4] as our feature extractor. This

(FPN based SSD) achieves state-of-the-art performance for a single model on PASCAL VOC dataset (object detection) as shown here[1].

On the Target dataset, many of the target classes cannot be distinguished by visual features alone. Hence we resorted to using a separate classifier (as opposed to the FRCNN based LSTD classifier) for the detected boxes. This involves taking text based features. Hence, while fine-tuning, a better alternative to this classifier is used in our system. The learning of target domain is made easier and faster by making use of the background regularization constraint.

4 Methodology

The task can be described as few shot document layout understanding. Our methodology consists of the following parts

1. Creating the artificial (Source) dataset.
2. Pretraining the model on the Source dataset.
3. Finetuning the model on the domain-specific (Target) dataset.
4. Training the ML classifier on the Target dataset (is combined with Step 3).

4.1 Dataset Generation

Our artifical dataset contains 160,000 images spanning multiple scales and sizes, accommodating for asymmetrically placed structures and elements. The dataset contained 8 basic layout classes: Title, Heading. Sub-Heading, Text Block, List, Table, Image Content, Image/Table Caption.

The textual content in the dataset was taken from a text dump consisting of a variety of online sources. The images were taken from a small dataset collected from Google Images. Apart from random images, the image dataset contained specific images collected using relevant keywords like graphs, tables, charts etc.

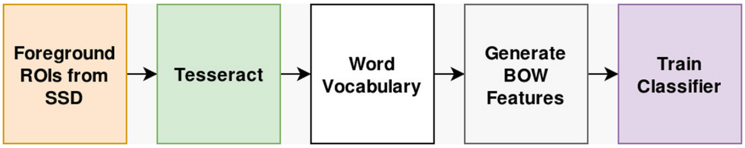


Fig. 2. Overview of the ML Classifier

4.2 Training

We train the LSTD model as it is on the Source Dataset. Once our model is trained on the Source Dataset, we move to fine-tune the model on the Target Datasets. Here we apply BGR. As mentioned earlier, we found that the performance of the inbuilt classifier in LSTD was not performing to our satisfaction,

[1] https://github.com/kuangliu/torchcv.

hence we decided to pass the foreground detections from the network through a seperate classifier.

Target Classification: To tackle the domain specific layout classes, we employed few ways to extract the best features so that we can train a classifier. We extracted the text from the detected box and used bag-of-words approach for getting the textual features. We also used other features related to the spatial configuration of the detected box. We use these features to train a machine learning algorithm to classify the detected bounding box to one of the classes. This is described in Fig. 2.

4.3 Implementation Details

For creating the artificial dataset, we generated HTML files which correspond to web documents and exported them into images using a webdriver. For the layout detection step, we implemented the LSTD network in PyTorch library. We use the FPNSSD from torchcv library (see footnote 1). For all experiments, we use SGD optimizer with learning rate of 0.0001 and momentum 0.9. We use L2 penalty of 0.0005. For the **layout classification** step, to extract text from a detected box we use the open-source LSTM-based Tesseract 4.0. We get our classifier using the tpot toolkit [9], which uses genetic programming to optimize machine learning pipelines. While reporting the results, we take the IoU threshold for evaluating object detection metrics as 0.5.

5 Invoice Dataset

We collected 170 invoices which includes variations in structure, domain and template. We refer to this as the Invoice Dataset. We manually tag this dataset into layouts of 5 main categories: Logo, Address, Bill/Invoice Information, Tables, Amount Information (Total). We use a fixed set of 100 images as our test set. We train our model on different (incremental) number of training images (k) and report the results correspondingly.

Table 1. LSTD end to end performance on **Invoice Dataset**

No. of training images (k)	Mean precision	Mean recall	Mean F1 score
10	0.4721	0.5188	0.4943
20	0.4962	0.5444	0.5192
30	0.5012	0.5791	0.5373
40	0.5244	0.601	0.5601
50	0.5316	0.6101	0.5682
60	0.5599	0.6214	0.589
70	0.56	0.6354	0.5953

6 Resume Dataset

The resume dataset is a set of 100 images collected from various sources containing resumes from different domains and layouts. As with the invoice dataset, this was manually tagged into 6 main categories: Education, Experience, Bio, Skills, Summary, Other. A fixed set of 50 images is used as the test set and training is done on an incremental number of training images ranging from 10 to 50.

Table 2. LSTD End to End performance on **Resume Dataset**

No. of training images (k)	Mean precision	Mean recall	Mean F1 score
10	0.6144	0.5888	0.6013
20	0.6398	0.6011	0.6198
30	0.6587	0.6218	0.6397
40	0.6712	0.6325	0.6513
50	0.6946	0.634	0.6629

7 Baselines

Table 3. Baseline (Docstrum) performance

Dataset	Precision	Recall	F1 score
Invoice	0.0547	0.1935	**0.0853**
Resume	0.2415	0.2559	**0.2485**

The Docstrum algorithm [8] serves as our baseline. The algorithm finds the connected components and their centroids. It then looks for the K-nearest neighbours (K = 5) of each component. Vectors are plotted from each centroid to its neighbours and these angles help in skew correction. The nearest-neighbor distance histogram has several peaks and these peaks typically represent between-character spacing, between-word spacing and between-line spacing. These values are then used to construct lines, words and text blocks with some predetermined tolerance for each spacing value.

We use Docstrum to construct blocks and then evaluate the outputs using the manually annotated ground truth boxes on both the target datasets ie. Invoices and Resumes. The results are reported in Table 3 while sample outputs of the method are shown in Fig. 3. Results of Docstrum can be compared with the foreground detection results (Table 4, Table 5) as end to end layout detection uses textual features.

8 Results

We perform multiple experiments to evaluate the performance of the proposed approach. We first show the effectiveness of source pretraining (SP) for few shot layout detection. Later, we evaluate our pipeline in 3 ways.

1. Evaluation of Foreground detection task
2. Evaluation of ML Classifier
3. Evaluation of end to end layout detection task

We evaluate the foreground detection performance of two types of models. In Tables 4, 5, *scratch* refers to the model which was trained from scratch, while SP refers to the model which was finetuned from the Source Pretraining (Step 2 in Sect. 4). One can notice an improvement of at least **40%** on F1 scores of Target Domain Layout Detection task. This signifies the importance of Source Pretraining in our proposed pipeline.

The evaluation of the ML Classifier on the foreground ROIs is shown in Table 6. The performance of our pipeline on the end to end layout detection task is shown in Table 1, 2 for the Invoice and Resume datasets respectively. The end to end pipeline consists of both the foreground detection and ML Classifier. We are able to obtain satisfactory performance even with 10 training images.

Table 4. LSTD foreground detection performance on **Invoice Dataset**. SP denotes Source Pretraining.

No. of training images	Precision		Recall		F1 Score	
	SP	Scratch	SP	Scratch	SP	Scratch
0	0.144	NA	0.4214	NA	0.2147	NA
10	0.5992	0.1078	0.6212	0.1991	**0.61**	0.1399
20	0.611	0.1377	0.7062	0.235	0.655	0.1736
30	0.6203	0.1744	0.7755	0.2768	0.6893	0.214
40	0.6767	0.1957	0.7901	0.2998	0.729	0.2368
50	0.6742	0.3018	0.7992	0.3036	0.7314	0.3027
60	0.7017	0.3738	0.8001	0.315	0.7484	0.3419
70	0.7292	0.3888	0.8132	0.3445	**0.7689**	0.3653

Table 5. LSTD foreground detection performance on **Resume Dataset**. SP denotes Source Pretraining.

No. of training images	Precision		Recall		F1 Score	
	SP	Scratch	SP	Scratch	SP	Scratch
0	0.035	NA	0.4311	NA	0.06	NA
10	0.8228	0.3797	0.821	0.3571	**0.8219**	0.368
20	0.8542	0.3859	0.8224	0.3928	0.838	0.3893
30	0.8655	0.5238	0.8291	0.5238	0.8469	0.5238
40	0.9123	0.5178	0.8363	0.7532	0.8726	0.6137
50	0.8977	0.6094	0.8343	0.61309	**0.8659**	0.6103

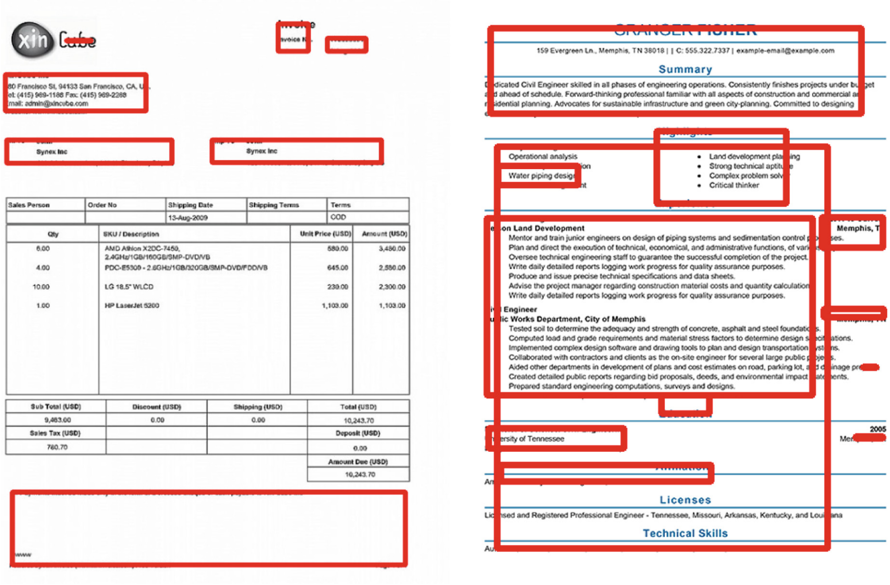
Table 6. Evaluation of ML Classifier on **Invoice** and **Resume** Datasets

Dataset	No. of training images (k)	Precision	Recall	F1 score
Invoice	70	0.7718	0.8135	0.7921
Resume	50	0.804	0.8946	0.8469

9 Conclusion

In this work, we have shown that object detection techniques can be used for Document Layout understanding. We have also shown that the proposed methodology can be scaled across multiple domains with just need of few tagged examples. The results also demonstrate the superiority of the methodology over existing object detection techniques. Document Layout analysis techniques assumes great importance in the information age as more and more documents are digitized and needs to be retrieved by understanding their content similar to digital content. Such techniques are useful in automating manually intensive business processes such as processing KYC documents or invoices. Document Layout analysis techniques also opens up the possibilities for businesses to mine documents such as paper receipts and extract valuable insights from them for market research purposes. Getting a large annotated corpus of data can be time-consuming and expensive for practical use-cases which further demonstrates the practical utility of our approach.

10 Qualitative Outputs



xin Cube

nvoice N .

80 Francisco St, 94133 San Francisco, CA, U .
el: (415) 969-1188 Fax: (415) 969-2288
mail: admin@xincube.com

Synex Inc

Synex Inc

Sales Person	Order No	Shipping Date	Shipping Terms	Terms
		13-Aug-2009		COD

Qty	SKU / Description	Unit Price (USD)	Amount (USD)
6.00	AMD Athlon X2DC-7450, 2.4GHz/1GB/160GB/SMP-DVD/VB	580.00	3,480.00
4.00	PDC-E5300 - 2.6GHz/1GB/320GB/SMP-DVD/FDD/VB	645.00	2,580.00
10.00	LG 18.5" WLCD	230.00	2,300.00
1.00	HP LaserJet 5200	1,103.00	1,103.00

Sub Total (USD)	Discount (USD)	Shipping (USD)	Total (USD)
9,463.00	0.00	0.00	10,243.70
Sales Tax (USD)			Deposit (USD)
760.70			0.00
			Amount Due (USD)
			10,243.70

www

159 Evergreen Ln., Memphis, TN 38018 | | C: 555.322.7337 | example-email@example.com

Summary

D dicated Civil Engineer skilled in all phases of engineering operations. Consistently finishes projects under bu get a d ahead of schedule. Forward-thinking professional familiar with all aspects of construction and commercial a r sidential planning. Advocates for sustainable infrastructure and green city-planning. Committed to designing

Operational analysis
on
Water piping desig
nt

- Land development pla ing
- Strong technical aptitu e
- Complex problem solv
- Critical thinker

e on Land Development Memphis, T
Mentor and train junior engineers on design of piping systems and sedimentation control p ses.
Plan and direct the execution of technical, economical, and administrative functions, of vari
Oversee technical engineering staff to guarantee the successful completion of the project.
Write daily detailed reports logging work progress for quality assurance purposes.
Produce and issue precise technical specifications and data sheets.
Advise the project manager regarding construction material costs and quantity calculation
Write daily detailed reports logging work progress for quality assurance purposes.

i l Engineer
u lic Works Department, City of Memphis
Tested soil to determine the adequacy and strength of concrete, asphalt and steel foundatio s.
Computed load and grade requirements and material stress factors to determine design s ci ations.
Implemented complex design software and drawing tools to plan and design transportation yst ns.
Collaborated with contractors and clients as the on-site engineer for several large public p oje s.
Aided other departments in development of plans and cost estimates on road, parking lot, a d d inage pr
Created detailed public reports regarding bid proposals, deeds, and environmental impact ate ents.
Prepared standard engineering computations, surveys and designs.

2005
ersity of Tennessee Me

An

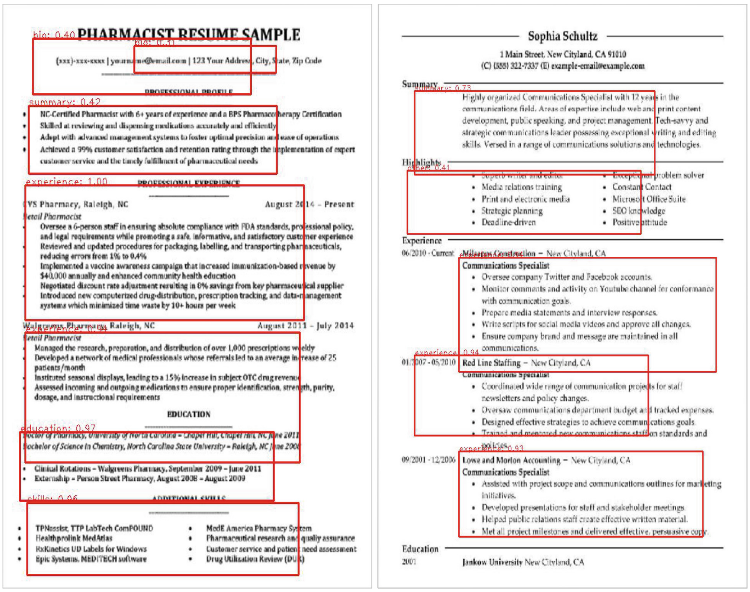
Licenses

Lic nsed and Registered Professional Engineer - Tennessee, Missouri, Arkansas, Kentucky, and Lou ana

Technical Skills

Au

Fig. 3. Sample predictions of the baseline method on both Datasets



bio: 0.40
bio: 0.51
summary: 0.42
experience: 1.00
experience: 0.94
education: 0.97
skills: 0.96
summary: 0.73
other: 0.41
experience: 0.94
experience: 0.94
experience: 0.93

PHARMACIST RESUME SAMPLE

(xxx)-xxx-xxxx | yourname@email.com | 123 Your Address, City, State, Zip Code

PROFESSIONAL PROFILE

- NC-Certified Pharmacist with 6+ years of experience and a BPS Pharmacotherapy Certification
- Skilled at reviewing and dispensing medications accurately and efficiently
- Adept with advanced management systems to foster optimal precision and ease of operations
- Achieved a 99% customer satisfaction and retention rating through the implementation of expert customer service and the timely fulfillment of pharmaceutical needs

PROFESSIONAL EXPERIENCE

VS Pharmacy, Raleigh, NC August 2014 – Present
etail Pharmacist
- Oversee a 6-person staff in ensuring absolute compliance with FDA standards, professional policy, and legal requirements while promoting a safe, informative, and satisfactory customer experience
- Reviewed and updated procedures for packaging, labelling, and transporting pharmaceuticals, reducing errors from 1% to 0.4%
- Implemented a vaccine awareness campaign that increased immunization-based revenue by $40,000 annually and enhanced community health education
- Negotiated discount rate adjustment resulting in 0% savings from key pharmaceutical supplier
- Introduced new computerized drug-distribution, prescription tracking, and data-management systems which minimized time waste by 10+ hours per week

Walgreens Pharmacy, Raleigh, NC August 2011 – July 2014
etail Pharmacist
- Managed the research, preparation, and distribution of over 1,000 prescriptions weekly
- Developed a network of medical professionals whose referrals led to an average increase of 25 patients/month
- Instituted seasonal displays, leading to a 15% increase in subject OTC drug revenue
- Assessed incoming and outgoing medications to ensure proper identification, strength, purity, dosage, and instructional requirements

EDUCATION

octor of Pharmacy, University of North Carolina – Chapel Hill, Chapel Hill, NC June 2011
achelor of Science in Chemistry, North Carolina State University – Raleigh, NC June 2006

- Clinical Rotations – Walgreens Pharmacy, September 2009 – June 2011
- Externship – Person Street Pharmacy, August 2008 – August 2009

ADDITIONAL SKILLS

- TPNassist, TTP LabTech ComPOUND
- Healthprolink MedAtlas
- RxKinetics UD Labels for Windows
- Epic Systems, MEDITECH software
- MedE America Pharmacy System
- Pharmaceutical research and qualiy assurance
- Customer service and patient need assessment
- Drug Utilization Review (DUR)

Sophia Schultz

1 Main Street, New Cityland, CA 91010
(C) (555) 322-7337 (E) example-email@example.com

Summary

Highly organized Communications Specialist with 12 years in the communications field. Areas of expertise include web and print content development, public speaking, and project management. Tech-savvy and strategic communications leader possessing exceptional writing and editing skills. Versed in a range of communications solutions and technologies.

Highlights

- Superb writer and editor
- Media relations training
- Print and electronic media
- Strategic planning
- Deadline-driven
- Exceptional problem solver
- Constant Contact
- Microsoft Office Suite
- SEO knowledge
- Positive attitude

Experience

06/2010 - Current Milsaps Construction – New Cityland, CA
Communications Specialist
- Oversee company Twitter and Facebook accounts.
- Monitor comments and activity on Youtube channel for conformance with communication goals.
- Prepare media statements and interview responses.
- Write scripts for social media videos and approve all changes.
- Ensure company brand and message are maintained in all communications.

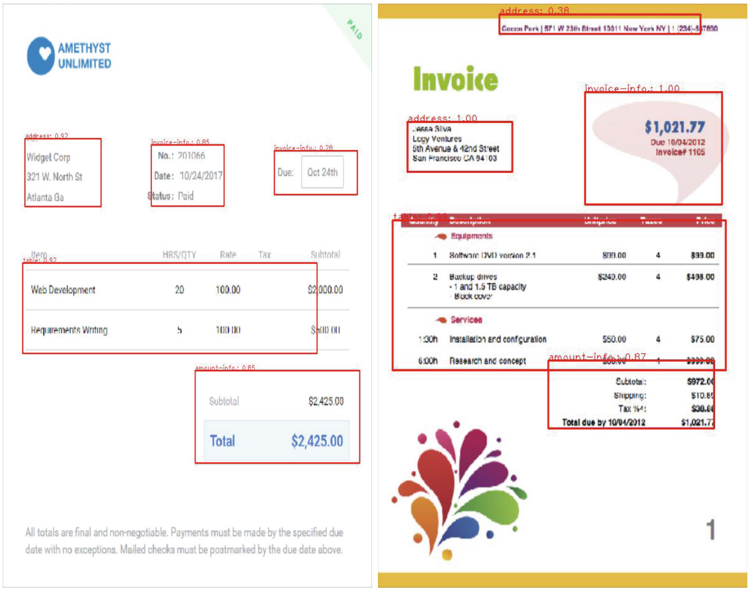
01/2007 - 05/2010 Red Line Staffing – New Cityland, CA
Communications Specialist
- Coordinated wide range of communication projects for staff newsletters and policy changes.
- Oversaw communications department budget and tracked expenses.
- Designed effective strategies to achieve communications goals.
- Trained and mentored new communications staff on standards and policies.

09/2001 - 12/2006 Lowe and Morton Accounting – New Cityland, CA
Communications Specialist
- Assisted with project scope and communications outlines for marketing initiatives.
- Developed presentations for staff and stakeholder meetings.
- Helped public relations staff create effective written material.
- Met all project milestones and delivered effective, persuasive copy.

Education

2001 Jankow University New Cityland, CA

Fig. 4. Sample predictions from our system on the test images of Resume Dataset


address: 0.92
invoice-info.: 0.85
invoice-info.: 0.28
table: 0.92
amount-info.: 0.85
address: 0.38
invoice-info.: 1.00
address: 1.00
table: 0.99
amount-info.: 0.87

AMETHYST UNLIMITED

PAID

Widget Corp
321 W. North St
Atlanta Ga

No.: 201066
Date: 10/24/2017
Status: Paid

Due: Oct 24th

Item	HRS/QTY	Rate	Tax	Subtotal
Web Development	20	100.00		$2,000.00
Requirements Writing	5	100.00		$500.00

Subtotal	$2,425.00
Total	$2,425.00

All totals are final and non-negotiable. Payments must be made by the specified due date with no exceptions. Mailed checks must be postmarked by the due date above.

Cocoa Park | 571 W 23th Street 10011 New York NY | 1 (234)-567800

Invoice

Jessa Silva
Logy Ventures
5th Avenue & 42nd Street
San Francisco CA 94103

$1,021.77
Due 10/04/2012
Invoice# 1105

Quantity	Description	Unitprice	Taxes	Price
	Equipments			
1	Software DVD version 2.1	$99.00	4	$99.00
2	Backup drives - 1 and 1.5 TB capacity - Black cover	$249.00	4	$498.00
	Services			
1:30h	Installation and configuration	$50.00	4	$75.00
6:00h	Research and concept	[illegible]	4	[illegible]

Subtotal:	$972.00
Shipping:	$10.85
Tax %4:	$38.88
Total due by 10/04/2012	$1,021.77

1

Fig. 5. Sample predictions from our system on the test images of Invoice Dataset

References

1. Agrawal, M., Doermann, D.S.: Voronoi++: a dynamic page segmentation approach based on voronoi and docstrum features. In: 10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26–29 July 2009, pp. 1011–1015 (2009). https://doi.org/10.1109/ICDAR.2009.270
2. Chen, H., Wang, Y., Wang, G., Qiao, Y.: LSTD: a low-shot transfer detector for object detection. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, 2–7 February 2018 (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16778
3. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval, vol. abs/1502.07058 (2015). http://arxiv.org/abs/1502.07058
4. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection, vol. abs/1612.03144 (2016). http://arxiv.org/abs/1612.03144
5. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
6. Mao, S., Kanungo, T.: Empirical performance evaluation methodology and its application to page segmentation algorithms. **23**, 242–256 (2001). https://doi.org/10.1109/34.910877
7. Namboodiri, A.M., Jain, A.K.: Document structure and layout analysis. In: Chaudhuri, B.B. (ed.) Digital Document Processing: Major Directions and Recent Advances. ACVPR, pp. 29–48. Springer, London (2007). https://doi.org/10.1007/978-1-84628-726-8_2

8. O'Gorman, L.: The document spectrum for page layout analysis. **15**, 1162–1173 (1993). https://doi.org/10.1109/34.244677
9. Olson, R.S., Bartley, N., Urbanowicz, R.J., Moore, J.H.: Evaluation of a tree-based pipeline optimization tool for automating data science. In: Proceedings of the 2016 on Genetic and Evolutionary Computation Conference, Denver, CO, USA, 20–24 July 2016, pp. 485–492 (2016). https://doi.org/10.1145/2908812.2908918
10. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks, vol. abs/1506.01497 (2015). http://arxiv.org/abs/1506.01497
11. Shafait, F., Keysers, D., Breuel, T.M.: Performance comparison of six algorithms for page segmentation. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 368–379. Springer, Heidelberg (2006). https://doi.org/10.1007/11669487_33
12. Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., Giles, C.L.: Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 4342–4351 (2017). https://doi.org/10.1109/CVPR.2017.462