DeepPDF: A Deep Learning Approach to Analyzing PDFs

Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells

Oak Ridge National Laboratory

Oak Ridge, TN, USA

{stahlcg, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov

Abstract

Scientific publications contain a plethora of important information, not only for researchers but also for their managers and institutions. Many researchers try to collect and extract this information in large enough quantities that it requires machine automation. But because publications were historically intended for print and not machine consumption, the digital document formats used today (primarily PDF) have created many hurdles for information extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for machine processing. While a number of tools exist, which can extract the contents of a PDF with acceptable accuracy, correctly labeling and piecing this data back together in a machine readable format is a significantly harder task. In this paper we explore the feasibility of treating these PDF documents as images, we believe that by using deep learning and image analysis it may be possible to create more accurate tools for extracting information from PDF documents than those that currently exist.

Keywords: deep learning, information extraction, scholarly publications

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan¹.

1. Introduction

Scientific publications have a wealth of information that can be useful for research, making management decisions, evaluating impact, etc. But often times this data is currently locked behind the PDF standard. While tools do exist to extract text and other information from PDF documents, the resulting output often falls short of the demands of researches. Captions, figures, tables, header, and footer data are among some of the features that cause problems for traditional PDF extraction methods, as well as for tools for analyzing extracted information built on top of these methods. For example, (Beel et al., 2013) have reported that while many tools claim around 90% accuracy on title extraction, their experiments with existing tools have resulted in accuracies between 50% to 70%. Furthermore, (Lipinski et al., 2013) have conducted a comparative evaluation of several tools including GROBID, which has revealed poorer performance on abstract extraction.

In this publication we explore the possibility of creating PDF analysis tools that treat PDF documents as images. We believe that scientific publications have inherent structure that is easy for humans familiar with them to separate.

¹http://energy.gov/downloads/ doe-public-access-plan

When presented with an image of a fully redacted publication researchers can visually determine the difference between a title, paragraph, reference section, headers, etc. Using the idea that PDF structure is a trainable idea, we theorize that a deep learning network can also be trained to separate the different sections of publications. Being able to separate out the different sections of publications is important because it will allow tools to accurately provide raw text versions of individual sections of the document, without the noise created by traditional methods. To test this idea, we have utilized a set of 50 publications from PubMed (a subset of the PMC_Sample_1943 dataset (Constantin et al., 2013)), which we have manually annotated (Section 3.), resulting in 407 labeled pages. We have then trained a Deep Neural Network to identify body text in the input (Section 4.). Our evaluation shows this approach offers high accuracy in correctly identifying body text while correctly rejecting other elements (Section 5.).

2. Related Work

In this section, we review previous literature relevant to our study, which we categorize according to the method used. First, we discuss methods for automated extraction of information from research articles which use traditional machine learning models such as Conditional Random Fields. Finally, we focus on methods which have utilized Deep Learning for related task.

2.1. Traditional methods

The analysis of the structure of documents has been studied for a number of years ((Mao et al., 2003) have provided a survey of some earlier approaches) and a number of freely available tools currently exist which can be used to extract information from scientific documents. These include ParsCit² (Councill et al., 2008), GROBID³ (Lopez, 2009), CERMINE⁴ (Tkaczyk et al., 2014), and most re-

²http://parscit.comp.nus.edu.sg/

³cloud.science-miner.com/grobid/

⁴cermine.ceon.pl/

cently OCR++⁵ (Singh et al., 2016). While previous approaches utilized models such as Hidden Markov Models (HMM) and Support Vector Machines (SVM) (Peng and McCallum, 2004), most of the current tools, such as ParsCit (Councill et al., 2008) and GROBID (Lopez, 2009), utilize Conditional Random Fields (CRF). CRFs are undirected graphical models trained to maximize a conditional probability (Peng and McCallum, 2004) which can be used to segment and label sequence data (Lafferty et al., 2001).

For example, ParsCit uses a CRF model to process reference strings to identify parts such as author, title, and venue information (Councill et al., 2008). The authors have used several heuristics to identify the reference section and to split the section into separate references. The CRF model is then applied to the separate reference strings. The authors also use heuristics and regular expressions to extract the context in which each reference was mentioned in the text. Following the approach of (Peng and McCallum, 2004), GROBID used CRFs for both header and reference string parsing (Lopez, 2009). CERMINE combines several models, mainly SVM which is used to identify zones (header, body, references, other) in the input text, and CRF which is used for parsing reference strings (Tkaczyk et al., 2014). OCR++ also uses CRFs (Singh et al., 2016). The tool uses several separate CRF models and combines them with handwritten heuristics.

2.2. Deep Learning methods

In recent years, neural networks have become the state-ofthe-art in a variety of computer vision tasks (LeCun et al., 2015). These networks consist of neurons arranged in series of layers, which learn to recognize successively higherlevel representations. To the best of our knowledge, only one previous study has treated PDF documents as images and leveraged Deep Learning for this task (Siegel et al., 2018). They have utilized a modified version of the *ResNet-101* network to extract figures and captions from scientific documents. In contrast to this work, we currently focus on body text identification.

3. Data Collection

The *PMC_Sample_1943* dataset compiled by Alexandru Constantine was selected for this project⁶ (Constantin et al., 2013). This dataset consists of 1943 publications selected from 1943 different journals in the Pubmed repository. For the initial testing we selected a random sample of 50 documents, giving us a total of 407 pages of publication data. Each section of the publication was assigned an RGB color code then using Adobe Acrobat's redaction tool, a researcher manually applied redactions to each section of the PDF documents. A copy of the PDF was then saved giving us a redacted or masked version of the document and the original document itself. The original document is converted to a grayscale PNG, while the masked document is converted to an RGB PNG. While this process for redaction is highly accurate, it is manually intensive. We believe

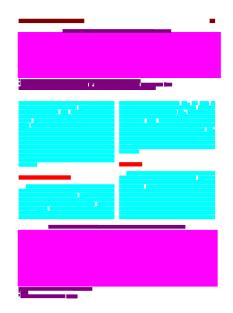


Figure 1: Example of redacted document where redaction color indicates the type of content for each pixel.

that while there are many different journals and publication venues, scientific publications cluster into a much smaller number of visual differences. Because of this we believe that a smaller dataset can be used for proof of concept and that results can be achieved with a much smaller data set than was is generally required for image analysis problems.

4. Methods

As the data has been processed such that we have an image of each page of each PDF and a pixel-wise label for the type of content corresponding to each pixel, the problem is naturally set up for semantic segmentation. Semantic segmentation is the process of assigning a label to each pixel of an image. A popular network for semantic segmentation tasks is *U-Net* (Ronneberger et al., 2015), and this is the network architecture we chose to utilize in this work. This network was chosen as it typically provides good performance with relatively few training examples.

We used a *U-Net* implementation available at https: //github.com/shreyaspadhy/UNet-Zoo and the network is trained using softmax cross entropy loss. For this paper, we are only exploring a two class problem where one class is "paragraphs" and the other is "not paragraphs." The former is defined as the main text of the paper, and the latter includes titles, authors, author information, blank space, figures, tables, references, abstracts, etc. We target this problem since existing PDF text extraction tools often fail at separating the main text from these other text within the document, making the result challenging to read (Section 1.).

5. Results

In order to evaluate this method, we split the 407 pages of publications in to 366 training examples and 41 validation examples. Figures 2 and 3 provide validation exam-

⁵www.cnergres.iitkgp.ac.in/OCR++/home/

⁶https://grobid.s3.amazonaws.com/PMC_ sample_1943.zip

	Migration of monomer stymme from GPPS and MIPS 228				
Samples were-purchased from markets and stored at CC	Table 1. Contents of segments of styrest constant from CPP1 raps				
until analyzed. The samples were usually stated for no larger	[ag 2] has been to define the president and they periods.				
then it days before being analyzed. The preparations of som- ples over prepared by the same procedure and by Raziak	Temperature (n) III III III				
and Sharmod Eriskii. The chromaingraphy was carried out	20 0.0 KG11KM 1341KM				
and pank heights were plotted applied streets concentra-					
tice. By estrapolation, the amount of styrene in the sample	NO TRAINE PROVIDENT ADDRESS				
was obtained. Calculation of the styrene concentration was					
by/FFLC software-based as peak area of the chromatograms, of sayone abtained from the sample and calibration stand-	Table 2. Cantesis of signation of styrese ransomer from CPP1 raps				
and an if the posh arrow of the incertail standards.	(and the state of the second s				
The calculation of second straves was done according	Temperature (11) 10 10 10				
to the Galibeation curve method by addition (2.5, 5, 18, 38,					
56 erg./() of styres/working solution (50.6 erg.() is mediated) in a strengt free serve for all the details (will reative state land).					
to a stypene-free sample of the deteks/publication standard]. An identical volume of internal standard was added to the	300 TAL+LIN A.32+LOI A.31+3.30 Units are next to Chore an upber of Chare reployies.				
distillate and the surrow. The bistic and calibration stand-	CHE AT ROUGH STORY IN CONTROLS.				
and solutions were also analyzed in the same war as the	Table & Controls of adjuston of eigener carmoner from CPP1 raps.				
samples were prepared. The extension times for styrene and	(ag T is a but room with in different imagenetices and trace ports in The Cartal				
a credital styrate way approximately 4.5 and 6.5 miration, suspectively, in the local stimulants (avoir acid, ethanol,	Tenerosiae (10) 10 10				
respectively, is the lood stimulants Jacobic acid, ethanol, allow off, storeng was challed and dependent in the same	20 0.0 6711004 1.3119.00				
procedure. A known amount of styrene monomer (0.11 mg/					
kg) was added to some drink samples to determine recon-	100 AGE 105 BAS BAS AGE 100 Des av nexes chlore sample of these spheres.				
ery of styrme menoraer from different drinks. The limit of					
detection of the analytical method for styrene analysis in					
drink samples sas/1.081 mg/kg.	(ap 7) in a har one in different temperatures and time pretorie. There (min)				
	terestation (tr) of the fit				
Results and discussion	40 80 KM 1998 1481158				
One hundred and shire-two samples, including crostal and					
impact caps with hot different drinks such as important, and	Den av menn of fore an acte of these reductors.				
more will, were analyzed for determination of edgestion of strenge menomer, in all cases, relation occurred and					
the amount of structure monomer insected was in the came-	table 5. Convers of angestan of storms account from HTS raps	-			
af a contrained on? for crystel and impact cups with wa mile.	ing Q as a feer sells to defense compensations and these pottods.				
and cucus mills. Our finding of styreno detection with 10%C	Temperature (11) 18 28 19				
method is similar to fire mudy of Flanjsk and Sharrad [1864]					
when they used both DC-FID and HPLC methods. However, sur-limit of detection was lower (1.00) reg/lg/share/he HPLC	00 3.42+8.32 4.11+6.01 3.98+8.01 100 6.94+8.00 7.22+8.00 7.85+8.00				
[0.001 mg/kg] and GD-FED (0.01 mg/kg) methods of Flangek	ing an many of the range of these reductions	-			
and Sharnad (1994) when styrene-mores mer was determined					
in the lood packaging of dairy products. This may indicate					
that our IPLC method of extraction and detection is more sensitive than the GC FID and IPLC methods of Flanish and	pg (months and an anno regimes and an prove		-		
Sharan(1) 96-0. The concentrations of a percentation from	Proprietar (%) 10 Ki				
these samples are presented in Tables 1 in 6 with means and	20 0.0 0.72+0.00 1.14+0.07				
	00 AN1+0.02 4.02+0.00 4.02+0.00 V0 7/0+0.00 7/0+0.00 7/0+0.00				
to 0.12 year of the crystal rups with hot new, item 2.02 hot 30 years for crystal curs with hot mills, from 2.11 to 0.01 years for crystal	the an annual of these samples of these subjects.				
cars with her come mile, from 0.00 to 0.01 age for crystal					
rups with hot are, the point is to 7.65 up 11 for impact cups with	Table T. Recovery of added signed-monitors in ratiosa drink sa agles. Manuscript signed. Manuscript				
hot wilk, and from 0.72 in 7.78 up7 for impact cups with hot					
	Ten 0.00 0.007 ML7 MCB 0.00 0.0070 MC7				
Statistical analysis of results by ANOA's showed scriggelf- unst differences among all samples. Increasely of added sty-	Mik 0.0 82015 57.2 Generalik 0.0 8205 55.2				
more differences among all samples. Accounty of added sig- more menomer from various drinks is environed in Table 7.	Total and a second different second s				
	() Т	1 1	1	1 1	
	(9) F	Example of prope	erly rejecting ta	hles	
	(a) L	maniple of prope	my rejecting ta	UICS.	
			• • •		
Seture et al. Model insmal of Deergency Surgery 2011, #16					

<text><text><text><text><text><text></text></text></text></text></text></text>	The standard and standard	Kathura et al. Miteld Journal of Desegony Jorgery 2011, &16 Mign/Territor inforcerg/content/071116			
Image: A start of a star	The standard and standard	These seese no store of bound inchesting such as bound	condition false houses now the subscend in the mid to	 	
Image: state sta	Image: state stat				
Image: State Stat	Image: State Stat				
Image: A start of the star	Image: A start of the star				
Image: state in the state	Image: state in the state				
 A start of the sta	 A start of the sta				
Amount	Amount				
Air gamma and any damp any d	Air gamma and any damp any d				
The product of end of the product o	The product of end of the product o				
mail	mail				
Image: starting starti	Image: starting starti				
Kar and a set of the set	Kar and a set of the set				
Ale and a lange of the second	Ale and a lange of the second				
Image: state	Image: state				
Image: Section 1 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Imag	Image: Section 1 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Image: Section 2 Imag				
Image: state	Image: state	processing any a and the symptom five 3 years after	Construction areas areas of the Manadary		
In the second s	In the second s	reduct's some an arrestated electronic red growne han-	reportation of a state of the SMCA is a rary condition		
Image: A start of a star	Image: A start of a star	pression, 3 year after surgery, a thromboard false liamen	and is not associated with antic dissection. It was first		
		company second with sloer like projection (ULP) on	concreted by transitions in 1947 [19]. In previously		
Al constrained with a second of the second o	Al constrained with a second of the second o	follow up CT (figure 2b).	reported cases before 1972, the prognosis was very poor		
Al A set a la fair a la fa	Al A set a la fair a la fa		[19,50] However, the programs has improved signifi-		
A set data if a data if	A set data if a data if	Case 3	cantly since 1875 as a result of advancements in surgical		
Image: Section 1 and Section 2 and Sectio	Image: Section 1 and Section 2 and Sectio	A 47-year-old max with a 5-day history of south spi-	techniques and imaging modalities [1-4].		
We are the state of the sta	We are the state of the sta				

(b) Example of properly rejecting figure and caption.

Figure 2: (Left) Input image. (Center) Network output. (Right) Ground truth.

independent managing problement in pitched and the	C. Huden at al.
5. Balacillar, J. & Newley Reaction, A. (200). A strahoustic	membran in annual marker with the Grant Practice Assessment
mit: Sundracian Jamel of Psychology, 65, 521-529. 31. Marca, D. M., Ward, J. & Chandes, K. K. (2007) Averaging pro-	 Menor, E. W. & Howse, J. G. (2006) CQL2 - a new sension of Industry historycond care in estance care consultation. The Decid.
 Mill, A. T., LL X. M. & Thompson, D. K. (2005) The concerning solution of the University and the University Systems and Auto- 	ability and validity of a that various of the Oscial Platfiel Associa- ness Department (PDM), Associal of the Medical Association of
56. Indiana C., Styles, L. Watt, A. 56. & Indiana, M. (2006) The pay-	T2. Paulines, M. (2001) A measure of patient empowerment in longitud
domento: of devikying the patient empowerment radie. Accord of Princhesental Obserlags, 19, 1–15.	enversements cauching for elder people, increase of Advanced Naming, 34 (20), 486.
(201) PD Barrahas, Varenty of Constraint, N. Person, N.A. Laux, J.E. Harley, I. Erid B.J. Zonowan, M.	
 Peterson, N. A., Lawe, J. B., Hughey, L. Renk, R. J., Neuwernan, M. A., & Sares, F. W. (2005) Ubseuclos for Internet-real assessment of 	 Kennen, T., Litsuniers, L., Villerg, J. & Peter, U. (2000) Deut- mins restriction levels connecting measurement performance.
10. 201 (201 (201)) 20. 2010; A. T., Marris, C. R., Thompson, D. K. & Weiss, K. T. (2009)	councility - a case mich; carto patient concenter in a hospital.
 Linkerk, J., Panerk, M., Sanderg, G., Smidr, B., Wilsond, G. & Wildrich, K. (2007) Prochemotics mountain of the Sandhik Difference 	 Steaking, O.L. (2017) Using larve maryon for validity monomer- method contributions. Journal of Neurice Houseware, 11, 2015.
Emprovement Inde, Austinetics Instead of Carlog Science, 75.	
 Rosto, E., Markella, F., Degli Especial, M. & Moneste, F. (2007) Perchastratic properties of the 1025M, Judge sensing of the Tim- 	 A. Sawari, M. (2014) Effect on leading-school conceases of inner- ventions to give the Harrischer between patients and practitioners. a
Perdometric properties of the \$2550, Subar senses of the Yim- metric excellenced work to excessive exponenties among sum of	
citantic Socials, 55, 256-254. 62. Honier, 8 G., Henney, D. J. & Maxwell, 56, (1997) Measuring-public	Courses Exercisponer in Norving: Franksister, Techniques and Applications, Ind etic (etic R. L. Boulgers, B. K. A. Kowly, pp. 77-112).
	tion in Naming Tax Norvalk, CT: Appleton & Lange, 31 Mart 20 & Dones II (2007) Debug counted constitutions and on-
 Bowis, J. D., Banny, G. J., Marwell, M. & Walker, J. 2 (1997) A comparison of a Pathene trackness trackness PhD: space rev 	 Shad, N. & News, P. (2002) Palent-orange complations and on- comes in primary care : a service of the Decision. <i>Palent Education</i>.
 Pher, D. Haughery, J., Boren, J. P., Merrien, K. & Kenter, I (2007) Validation of a modeled mattern emblement instrument in automa 	Advanced Naving, 16, 558-560. 83. Compt. C. D. (1996). Associate client screatile: vitationi associated
Waldstim of a modeled patient emblement inclusion in action (absorpt). The Encause Restington Partial, 22 (1998) 471.	 Compte, C. D. (1994). Associate close scoregite: closed announces. Interclass: restrictioners, Junial Work, 74, 262, 208.
& Winniss, D. (2007) Massating empowerment among people with predictly directory a comparison of these technicals. Providence	From name crossed faulth counseling to empowermental localiticities willing. Patient Education and Counseling, 45, 69-78.
	Newbory Perk, Ch. Sage, 37, Th. YD. Mann, R. E. & Yeh, S. H. (2004). Relationship between
in preparaty the preparaty related responsements with (Alternet) formed of Walactics & Wanne's Hardle, (J. 191).	 Tu, Y. C., Nang, R. H. & Yen, S. H. (2000) Relationship leaves serviced researching and analysis of kin second eldels. Incl.
67. Bowie, I. O., Humpy, D.J., Marwell, M., Walker, H.F.& Persona, G.	denis will de saming factors in Talwars a garafamater samey daar.
 (2005) Developing a transalation quality index? (CQD) for one in second structure. Funds. Proceedin, 15, 473–495. 	national Assessed of Marcing Danlars, 43, 473, 680. 85, 305 Key, Ku Perfers, J. A. B. Braumen, K. (1980). Engineering in
provid practice Fuency Practice, 17, 403–405.18. Kim, S. C., Hunne D & Science, S. L. (2011) The Kim Allocov Institu-	 Shikkey, K., Polan, J. A. & Bourses, K. (1980) Expression in available postlari On UEspin of ONTE. Astrophys. Family Physician.
114.111. 11. Charles, C., Julieser, S. & Mardels, S. (2017) Interveting for patient	 Tatley, B. (2007) Experience of the propiet. Having (Killer People, 19, 12–16.
Pr. Caller, C., Salakie, C. & Handal, K. (2001) algebras and party	
1996	
5.92	

(a) Example of properly rejecting references.

[19] K. Myadai, K. W. Hashim, H. Matsuhimu et al., "Proble and chick depillonce of and conduct strategies at bodies found in policies with optimary bidapy sireheads a strategi- erater study." <i>Journal of Autoinstanty</i> , vol. 23, no. 3, pp. 2471-364, 2005.	[11] M. B. Yang, I. H. Yu, A. Nabajian, D. Nacherg, K. Lindor, and D. D. Hold, "Do artitucine antibulity in prinney- biling vielesis patients theirig increased with the law follow," <i>Climid Concentrality on Physical Systems</i> , vol. 3, no. 33, pp. 108–1132, 2004.	 <u></u>	
[40] H. Nakaman, H. Eonley, T. Merk et al., "Anti-gg/UI and anti-controllary antibodies are different risk factors for the programion of primary biliary circlassic," <i>Hypotology</i> , vol. 45, 40	[55] H. E. Genhein and I. E. Madag, "Primary Milary circlosis patieliges or peridex for autoinsensity," <i>Garcierentology</i> , vol. 100, no. App. 823–835, 1901.		
 (a) L pp 110–127, 2005 (d) D. E. Doplinson, C. Linsken, A. Pares et al., "Anti-pp150 antibody mirrory disease severity in primary bilary circle. 	[17] M. J. Genheim and J. S. Mackay. "The causes of prinnery billary circlosic convenient and incommission trachs," <i>Physiology</i> , vol. 17, no. 3, pp. 217–213, 2008.		
rule," Hypotology, will 40, no. 6, p. 1105, 2003. [41] D. P. Euglieron, C. Liaskon, E. L. Rappendon, and G. N. Dadekon, "Next mitocheradekal antibodies in parisests with	[0] I. E. Maskey, S. Multiinghum, K. Fide et al., "The product activities multi-orporary biliary circlosis," <i>Immunological Reviews</i> , vol. 174, pp. 236–237, 2000.		
eventic lopus erythematicus: rending the uniterests," Clinic Chevin Arts, vol. 575, no. 1-2, pp. 145–184, 2006. [05] D. F. Duslanin, A. Farra, J. Endo, et al., "Primar: Miner-	[39] D. F. Bophnes and D. Yogani, "Origin of cross-ranche accounting in primary bilary carbonic," <i>European actional vol. 26</i> , no. 6, or 633–635, 2005.		
sirdenk specific animalase antibulies in patients from Spairs," American Jaccust of Gatwonienings, vol. 90, no. 4, pp. 253–250, 2004.	[00] D.F. Bagdama and D. Wegani, "Its teriorant primery blary circlesis," Clinical Ecolors in Allergy and Immunology, vol. Nume. 1, pp. 10–19, 2006.		
[44] D. P. Bogdono, D. Verpai, P. Montori, L. Matneri, and E. E. Mandel, "Specificity of anti-q-DO antibody for primery biling-circlosite," Standardinet Internet of Linearconteneing, vol. 20, no. 4, pp. 405–406, 2006.			
[41] F. Lowenick, M. Fulda, J. D. Enformation al., "International fee gates in making your completences are sub-fact of velocity action and scores: Ever disease in petraney billing detected," <i>Journal of Higosology</i> , vol. 54, no. 5, pp. 565–552, 2004. [46] S. Connell, K. Eiger, C. Soliti, et al., "A simulation hand assay for an interdeet which colorbadies: detyping amount of AMM.	[61] S. Warroda, M. Nalamino, H. Mikuda, E. Hoperbild, and Y. Silo, "ELA UEM IEE metrical immensionlanearies and managings of generative directogeneous complex in primary Dilary colridous relations of missional metricity in barron mechanismum dissues," <i>Journal of Experimental Medicine</i> , 01, 83, no. 5, pp. 1155–1085, 1985.		
regative primary bilary skylouis," <i>Miputelogy</i> , vol. 40, no. 3, pp.659–663, 2007. [47] D. Vergari and D. P. Beglanos, "Froktor meders in AMA-regulator PRC". Associate Journal of Conversion/logy, vol. 99, 1993.	[63] S. Shinoda, M. Sukamura, H. Shipuratin et al., "Meni-vey popular at human P2C 12 135 179 popular, distance- dominant T: cell spinop of princep bilary atoloxis," <i>Populating</i> , vol. 14, no. 5, pp. 3121–1218, 2008.		
mi, L. pp. 241–243, 2003. [46] J. N. Monoldi, H. C. Mitchelen, E. M. Dalmart, D. E. Janes, H. F. Barnendika, and D. G. W. Janas, "Datased bioney of only-prinney bilary clatheoid," <i>Enext.</i> , vol. 348, no. 4830, pp. 1099–1090, 1093.	[64] S. Shinode, J. Van Te Wene, A. Asawa et al., "Morniforation and process for trapping analysis of a constrom To-displayer result in anticultureful automatignum, in primary bilary circlessis," Journal of Clinical Intercipacion, vol. 202, no. 33, no. 1021 - 1240, 1940.		
[49] H. C. Mikhims, M. J. Barembins, and A. Dauldek. Traditive anticelucionatifial antibody has moved dilution phosphane in bilin primary biling checked. <i>Hypothys.</i> 103, 6, no. 6, pp. 1279–1396, 1996.	[61] D. S. S. Jones, "Davids of primary bilary sitebook," Got, vol. 565, no. 16, pp. 1613–1624, 2007.[64] C. Shiri, and M. E. Gardwein, "The role of environmental feature in primary bilary derivon," Threads in Immunology, rel. 53, no. 6, pp. 95–405, 2009.		
Loose, "Locklence and prevalence of primary lifting circlensis in the rate of Neurantic agent Type, England," International Journal of Epidemiology, vol. 20, no. 6, pp. 105–105, 1097. [11] E. S. Danni, C. F. Donni, A. L. Serverner, et al., "An Invite:	[67] C. Selard, P. Dowenkod, M. Zhain, M. Packle, M. F. Selalan, and H. E. Gordanin, "Ubran and controlousnasticy in printing bilary derivative Tensor and Journality, vol. 6, no. 7, pp. 101–104, 2005.	-	
to percent delexingence in primery biliny circlesic correlation with biosings? APARS, ed. 116, no. 5, pp. 804–602, 1998.101 E. D. Jacob, M. J. Gerdwein, R. Pencon, M. Fachen, N. Y.	[68] H. L.Q. McNally, S. Tholaro, and O. F. W. Janon, "Size transient receives mercial agents involved in the cause of primary billing circlessis". <i>Evolution: Invest space class: Charlot age analysis</i> ," <i>Opportunity</i> , vol. 30, no. 4, pp. 1004–1176, 2008.		
Empara, and E. J. Hardware, "Disary bilary distribution Dependings of 30 nm kpp 201–308 2008. [5] C. Swateddi H. H. Goldner, and H. Will, "Increase Incline against Vanilare dest," in primary Milary colorests, Sciences.	[60] D. Alsani, P. Sarawakai, P. Dawi at al., "Datagen warption in challenging transmit for programming of primary billing vields with "International Department," in <i>Proc. Sci.</i> , pp. 955–953, 2004. 128 K. Sanan, P. S. C. Saran, K. Barwar et al., "Chambian semi- rational systems, P. S. C. Saran, K. Barwar et al., "Chambian semi- rational systems, P. S. C. Saran, K. Barwar et al., "Chambian semi- temport of the systems, P. S. C. Saran, K. Barwar, S. Saran, S. Saran, S. S. Saran, S. Saran, S. Saran, S. S. Saran, Saran, Saran, Saran, S. Saran, Saran, Saran, Saran, Saran		
Against Tanisha and Star Panaey Handy Correlation, Construct Vol. 100, 117 (2017), 108	[20] K. Anano, P. K. C. Jong, K. Bage et al., "Orantial anno- bialist and mitochoof-the anometicus in perioasy Miker circlosic identification of arthbodies against a constron- menicomment, constraint, and food addition, 2 extraordi- add," <i>Internal of Internationg</i> , vol. 176, no. 9, pp. 3875–3883, 2005.		

(b) Example of not properly rejecting references.

Figure 3: (Left) Input image. (Center) Network output. (Right) Ground truth.

ples of an input provided to the network, examples of network output, and ground truth target output. These results demonstrate impressive results with such a small dataset. In particular, the network is able to reject header and footer text extremely reliably. The network rejects most abstracts, figure captions and references, confusing only some where the text formatting is extremely similar to typical paragraph text. The per pixel classification accuracy on the validation set was 94.32%, compared to a baseline of classifying each pixel as "not paragraph" which would provide 79.67% accuracy.

6. Conclusion and Future Work

In this paper we demonstrated that deep learning-based image analysis can be used to identify sections of scientific publications. Given the results from our current experiments, we feel that deep learning can be successfully used to enhance current PDF extraction methods, and based on our findings we plan to continue collecting data in order to further increase our networks results, as we feel many of the misclassified portions of text are due to insufficient training data that does not currently characterize features such as reference sections and abstracts sufficiently.

Our current results show that a deep learning network can successfully distinguish and learn the difference between the body text and other portions of a PDF document. The next step is to extend the approach to identifying each type of text (title, author, abstract, body text, etc.) rather than simply body text versus other. Additionally, we plan to increase the accuracy of our network by adding more data and to create an extraction tool that leverages the output of the deep learning network to extract text. While we are currently evaluating accuracy based on a per pixel count of estimated versus redacted image, an improved test of accuracy would be to leverage such an extraction tool to identify the per character accuracy of this text extraction approach.

7. Bibliographical References

- Beel, J., Langer, S., Genzmehr, M., and Müller, C. (2013). Docear's PDF Inspector: Title Extraction from PDF Files. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 443–444. ACM.
- Constantin, A., Pettifer, S., and Voronkov, A. (2013). Pdfx: fully-automated pdf-to-xml conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 177–180. ACM.
- Councill, I. G., Giles, C. L., and Kan, M.-Y. (2008). ParsCit: an Open-source CRF Reference String Parsing Package. In *LREC*, volume 8, pages 661–667.

- Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.
- Lipinski, M., Yao, K., Breitinger, C., Beel, J., and Gipp, B. (2013). Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*. ACM Press.
- Lopez, P. (2009). GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *International Conference* on Theory and Practice of Digital Libraries, pages 473– 474. Springer.
- Mao, S., Rosenfeld, A., and Kanungo, T. (2003). Document structure analysis algorithms: a literature survey. In Tapas Kanungo, et al., editors, *Document Recognition and Retrieval X*. SPIE, jan.
- Peng, F. and McCallum, A. (2004). Accurate information extraction from research papers using conditional random fields. In *HLT-NAACL 2004: Human Language Technology Conference of the North America Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, pages 329–336.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- Siegel, N., Lourie, N., Power, R., and Ammar, W. (2018). Extracting Scientific Figures with Distantly Supervised Neural Networks. In *To appear in ACM/IEEE Joint Conference on Digital Libraries in 2018 (JCDL 2018)*. ACM/IEEE.
- Singh, M., Barua, B., Palod, P., Garg, M., Satapathy, S., Bushi, S., Ayush, K., Rohith, K. S., Gamidi, T., Goyal, P., and Mukherjee, A. (2016). OCR++: A Robust Framework For Information Extraction from Scholarly Articles. *International Conference on Computational Linguistics (COLING)*, pages 3390–3400.
- Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M., and Bolikowski, L. (2014). Cermine – Automatic Extraction of Metadata and References from Scientific Literature. In *Document Analysis Systems (DAS)*, 2014 11th
- IAPR International Workshop on, pages 217-221. IEEE.