CS 6604: DIGITAL LIBRARIES — FINAL PRESENTATION
DEPT. OF COMPUTER SCIENCE, VIRGINIA TECH

# Classification and extraction of information from ETD Documents

BY JOHN AROMANDO, BIPASHA BANERJEE, BILL INGRAM,
PALAKH MIGNONNE JUDE, AND SAMPANNA KAHU

{JAROMANDO, BIPASHABANERJEE, WAINGRAM,
JPALAKHMIGNONNE, SAMPANNA}@VT.EDU

DECEMBER 5, 2020

# Acknowledgements


Dr. Edward A. Fox for his instruction and mentorship

Dr. Jian Wu at Old Dominion University for his guidance and advice on parsing and classification, and for providing us with annotated gold standard reference strings for us to use for testing and evaluating NeuralParsCit

Web Information Retrieval/Natural Language Processing Group (WING) at the National University of Singapore, for their work on Neural ParsCit and SciWING, and for sharing their processing scripts and training data

This project was made possible in part by the Institute of Museum and Library Services LG-37-19-0078-19

# Introduction


# Introduction

Millions of Electronic Theses and Dissertations (ETDs) are openly available online.

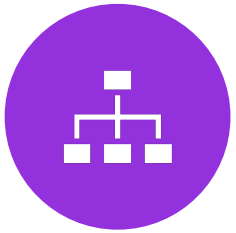
University Libraries VTechWorks repository holds over 30,000 ETDs, dating back to 1903 https://vtechworks.lib.vt.edu/.

The collection contains a mixture of born-digital and scanned documents.

Items contain PDF, Dublin Core metadata, and various supplementary files.

# Outline

INTRODUCTION, ETDS, UNIVERSITY LIBRARIES VTECHWORKS



CLASSIFICATION

REFERENCE PARSING

FIGURE EXTRACTION

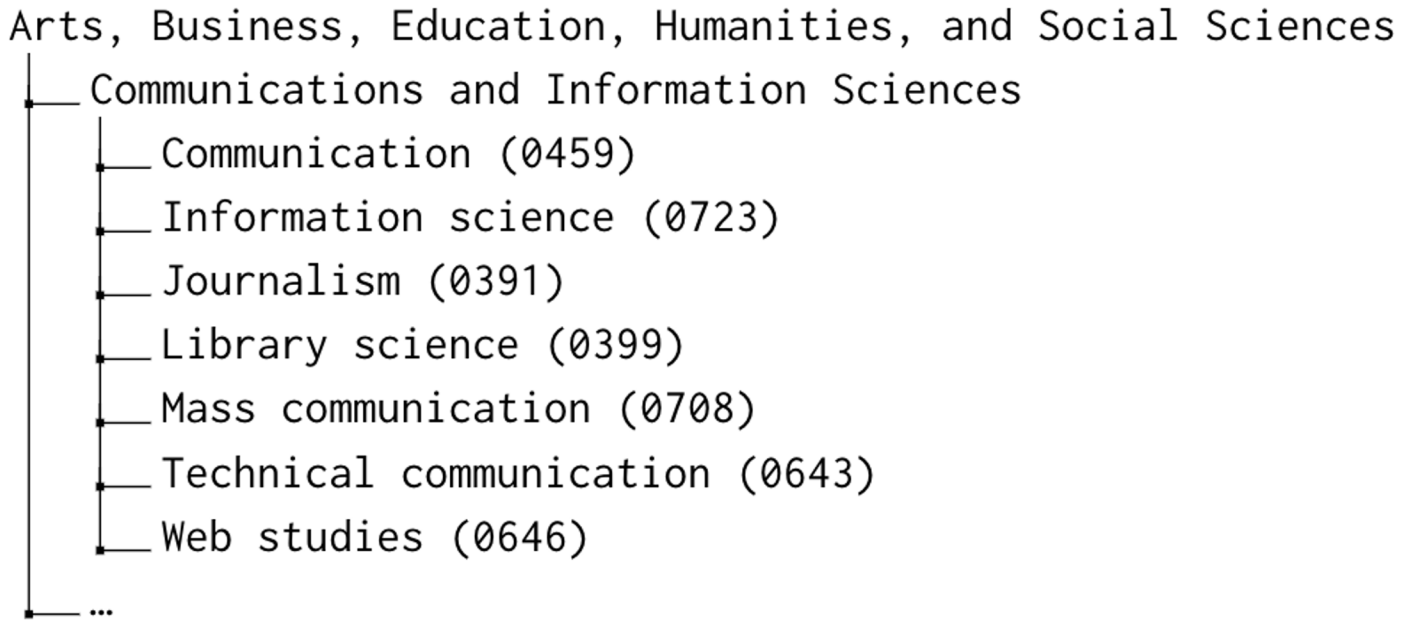
CONCLUSIONS AND FUTURE WORK

# Classification

# ETD Subject Classification

The **ProQuest Subject Categories** for the 2018--2019 Academic Year contain

- 2 primary headings
- 21 secondary headings
- 432 subject categories




Subsection of ProQuest Subject Categories

# Multi-label Classification

Authors are asked to pick one subject category that best describes their field of research or creative work and may choose additional categories as secondary subjects.

| Title | Primary subject category |
| --- | --- |
| *A history of the outplacement industry, 1960–1997: From job search counseling to career management. A new curriculum of adult learning* | Adult education (0516) |
| | **Secondary subject categories** |
| | Continuing education (0516) |
| | Management (0454) |
| | American history (0337) |

Example ETD with multiple subject categories

# Data Pre-processing

1. Formatting Classification Labels
2. Splitting Data
3. Data Cleaning
4. Dealing With Missing Values
5. Feature Selection
6. Evaluation Metrics

| PQ Metadata | VT Metadata |
|---|---|
| PQ_Author | VT_Subjects |
| PQ_Degree | VT_Title |
| PQ_Identifier_keyword | VT_Type |
| PQ_Number_of_pages | VT_contributor_author |
| PQ_Publication_year | VT_contributor_department |
| PQ_Title | VT_degree_level |
| | VT_degree_name |
| Derived_Advisor* | VT_description_abstract |

Set of features used to train the model

# Methods

1. Multi-Label Classification
   a. BinaryRelevance
   b. LabelPowerset
   c. Formatting class labels for the classification task

2. Machine Learning Algorithms
   a. Logistic Regression
   b. Support Vector Machine
   c. Random Forest

# Methods

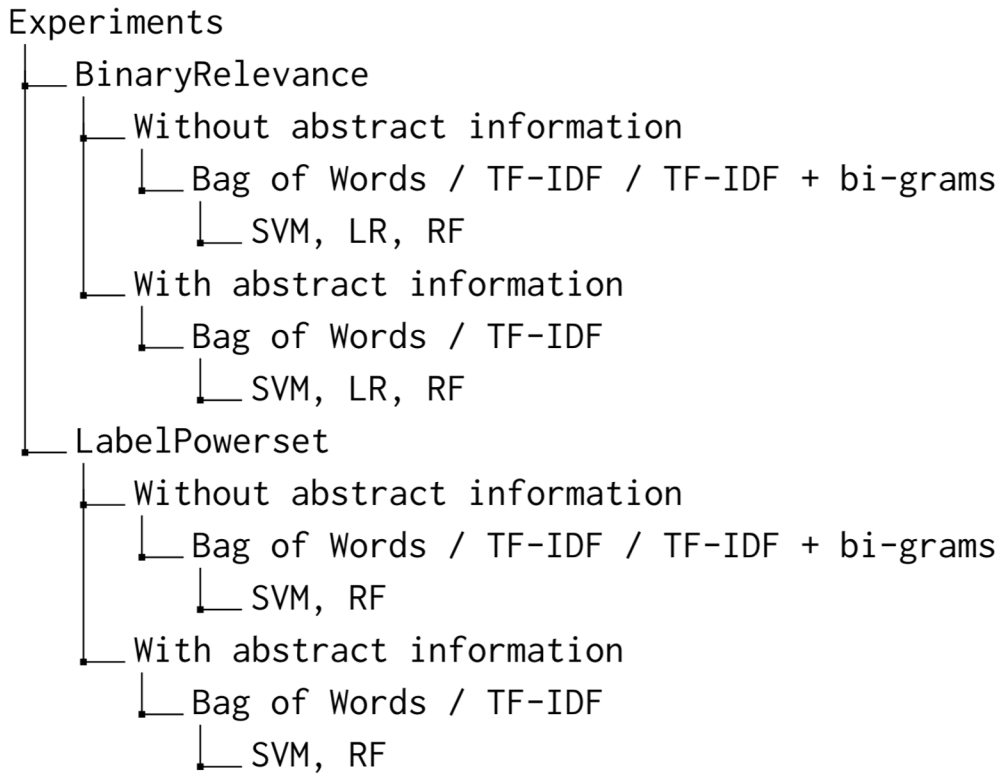
- Each category assigned a probability score
- 0 indicates not a label
- 1 indicates only class label
- Values between 0-1 indicate multiple class labels

$$Value = \frac{1}{number\ of\ classes}$$

| Statistics | Computer_science | Environmental_science | Forestry | PQ_Classification |
|---|---|---|---|---|
| 0.5 | 0.0 | 0.0 | 0.5 | Statistics, Forestry |
| 0.5 | 0.0 | 0.5 | 0.0 | Statistics, Environmental science |
| 1.0 | 0.0 | 0.0 | 0.0 | Statistics |
| 1.0 | 0.0 | 0.0 | 0.0 | Statistics |
| 1.0 | 0.0 | 0.0 | 0.0 | Statistics |

Sample ETD records with transformed class labels (with probability scores)

# Experiments




Various experimental setups

# Results and Discussion

Performance of various experimental setups using LabelPowerset with presence-absence labels (without abstract)

| Experiment setup | Accuracy | Precision | Recall | F1-score | Training Time |
|---|---|---|---|---|---|
| TF-IDF(bi) + SVM | 60.31% | 74.40% | 63.95% | 66.27% | 256.47 |
| TF-IDF(bi) + RF | 64.04% | 79.06% | 67.67% | 69.62% | 4.8 |
| TF-IDF + SVM | 61.29% | 75.62% | 65% | 67.55% | 97.91 |
| TF-IDF + RF | 57.17% | 73.50% | 60% | 63.54% | 0.82 |
| BOW + SVM | 58.93% | 73.92% | 62.45% | 64.43% | 85.47 |
| **BOW + RF** | **63.65%** | **80.41%** | **67.85%** | **70.70%** | **8.76** |

Performance of various experimental setups using LabelPowerset with presence-absence labels (with abstract)

| Experiment setup | Accuracy | Precision | Recall | F1-score | Training Time |
|---|---|---|---|---|---|
| TF-IDF + SVM | 61.88% | 76.41% | 65.41% | 68.05% | 169.24 |
| TF-IDF + RF | 60.11% | 77.96% | 64.14% | 66.21% | 15.12 |
| BOW + SVM | 44.79% | 64.03% | 49.37% | 51.27% | 162.74 |
| **BOW + RF** | **65.22%** | **79.64%** | **68.26%** | **70.37%** | **8.19** |

| PQ_Identifier_keyword | PQ_Title | VT_Subjects | VT_Title | VT_contributor_department | Expected_Labels | Predicted_Labels |
|---|---|---|---|---|---|---|
| applied sciences autonomous vehicles compressed sensing data compression haar wavelet transform occupancy grids slam simultaneous localization and mapping sparse signal reconstruction underwater autonomous vehicles | real time slam use compressed occupancy grid low cost autonomous underwater vehicle | autonomous vehicles slam occupancy grids haar wavelet transform compressed sensing sparse signal reconstruction data compression | real time slam use compressed occupancy grid low cost autonomous underwater vehicle | mechanical engineering | Computer_science | Electrical_engineering, Mechanical_engineering |
| health and environmental sciences communication and the arts social sciences business clothing disabilities innovations specialized product development women workers | innovation improvisation study specialized product development focus business clothing woman physical disability | working women clothing manufacturing co design universal design product development clothing physical disabilities | innovation improvisation study specialized product development focus business clothing woman physical disability | near environments | Marketing | Environmental_science |

# Results and Discussion

Performance of various experimental setups using LabelPowerset with probability scores (without abstract)

| Experiment setup | Precision | Recall | F1-score | Training Time |
|---|---|---|---|---|
| TF-IDF(bi) + SVM | 85.36% | 28.81% | 39.91% | 1030.83 |
| TF-IDF(bi) + RF | 87.16% | 30.25% | 43.13% | 4.01 |
| TF-IDF + SVM | 85.21% | 28.74% | 39.89% | 397.10 |
| TF-IDF + RF | 85.45% | 29.60% | 42.51% | 2.27 |
| BOW + SVM | 8813% | 29.60% | 41.74% | 352.63 |
| **BOW + RF** | **89.01%** | **30.77%** | **44.26%** | **2.26** |

Performance of various experimental setups using LabelPowerset with probability scores (with abstract)

| Experiment setup | Precision | Recall | F1-score | Training Time |
|---|---|---|---|---|
| **TF-IDF + SVM** | **84.69%** | **28.74%** | **39.93%** | **676.22** |
| TF-IDF + RF | 77.15% | 26.71% | 38.22% | 3.95 |
| BOW + SVM | 81.34% | 26.65% | 38.25% | 679.60 |
| BOW + RF | 78.76% | 26.78% | 38.64% | 2.89 |

| PQ_Identifier_keyword | PQ_Title | VT_Subjects | VT_Title | VT_contributor_department | Expected_Labels | Predicted_Labels |
|---|---|---|---|---|---|---|
| applied sciences education corgis computing comutational thinking data science datasets motivation | motivate introductory compute student pedagogical datasets | motivation introductory computing computational thinking engagement corgis datasets data science data pedagogical datasets computer science education engagement | motivate introductory compute student pedagogical datasets | computer science | Computer_science | Computer_science, Higher_education, Educational_psychology |
| biological sciences earth sciences decomposition hardwood forests microbial biomass nutrient cycling soil heterogeneity | soil resource heterogeneity site quality southern appalachian hardwood forest impact decompose stump geology salamander abundance | fine root dynamics appalachian hardwoods nutrient cycling soil heterogeneity ground penetrating radar salamanders decomposing stumps microbial biomass | soil resource heterogeneity site quality southern appalachian hardwood forest impact decompose stump geology salamander abundance | forestry | Forestry | Forestry, Ecology, Molecular_biology |
| social sciences brand equity customer equity hospitality businesses | create validate measure customer equity hospitality business link shareholder value return marketing | relationship equity brand equity value equity servperf firm valuation customer lifetime value shareholder value customer equity | create validate measure customer equity hospitality business link shareholder value return marketing | hospitality and tourism management | Management | Marketing, Management, Public_administration |

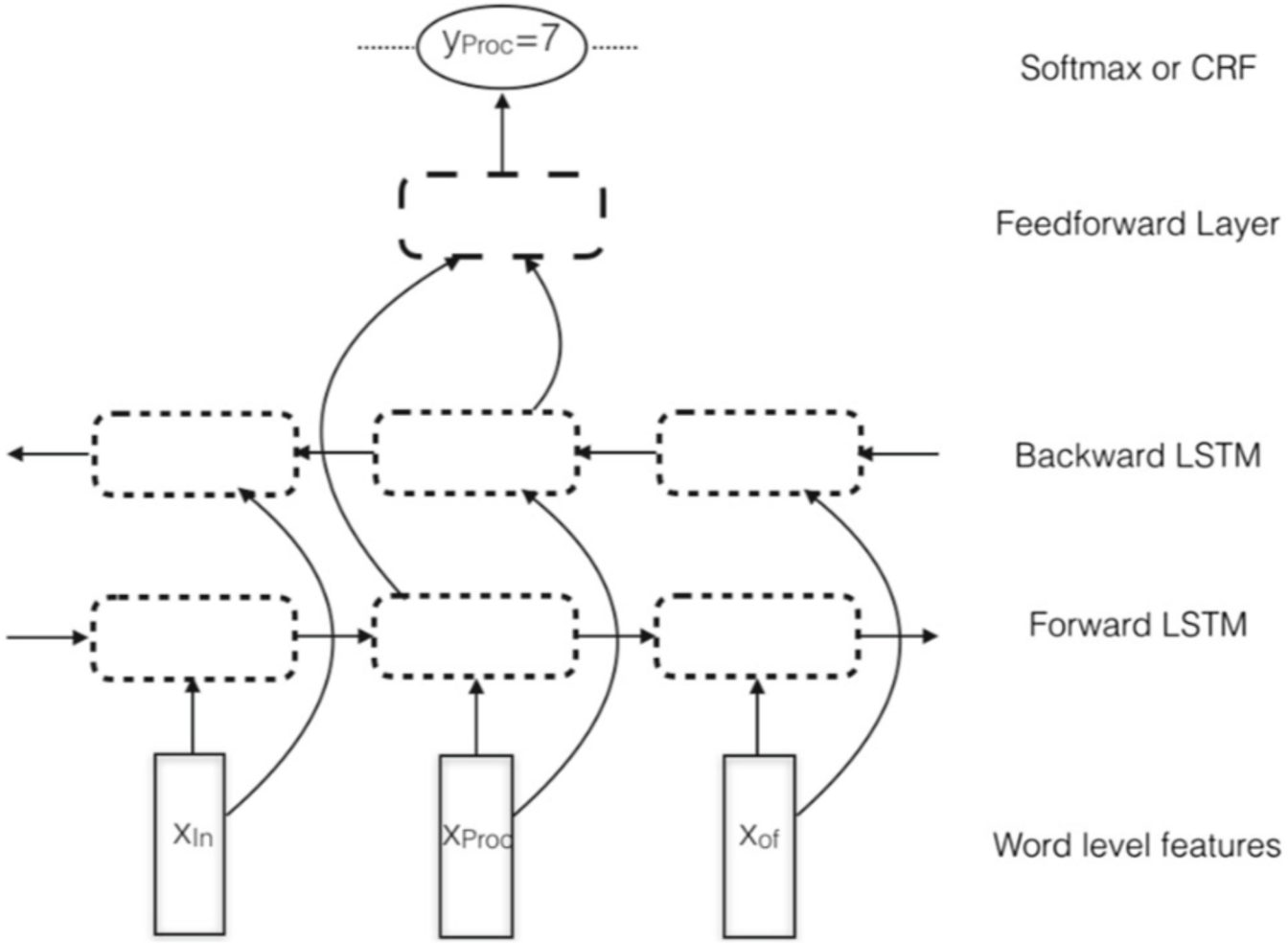
# Reference Parsing

# Neural ParsCit

Deep learning-based reference string parser

Developed by the Web Information Retrieval/Natural Language Processing Group (WING) at the National University of Singapore

Bidirectional LSTM




Neural ParsCit Model Architecture

Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018. Neural ParsCit: a deep learning-based reference string parser. *International Journal on Digital Libraries* 19, 4 (01 Nov 2018), 323–337. https://doi.org/10.1007/s00799-018-0242-1

# Methods

1. Download thousands of BibTex references from CrossRef REST API for multiple reference types (e.g., journal, book, conference paper)
2. Use Cite-Proc, a Citation Style Language (CSL) processor for Java, to generate multiple versions of each citation in 9 different citation types
3. Retrain Neural ParsCit with this data

| Citation style | Popularity |
|---|---|
| Modern language association (MLA) | 1 |
| APA | 2 |
| Chicago annotated bibliography | 3 |
| American medical association | 4 |
| American chemical society | 5 |
| National library of medicine grant proposals | 6 |
| IEEE | 7 |
| Turabian fullnote bibliography | 8 |
| Vancouver | 9 |

2015. Citation Styles Guide | Which Citation Style Should You Use? Scribbr. Retrieved December 3, 2019 from https://www.scribbr.com/citing-sources/citation-styles/

# Creating the training data

```
[2378] citation-number
R. author
Xu author
B. author
Aotegen author
and author
Z. author
Zhong author
, extra
“Synthesis, title
characterization title
and title
biological title
activity title
of title
C title
6 title
```

Training data format

1. Only conference proceeding and journal citations considered
2. Splitting data for each citation style
3. Combining data from each styles into one train, val and test sets

1. Shuffling the data
2. Create folds for cross-validation
3. Script to convert into the required format

# Training Process

1. Trained on ARC's Cascades cluster
2. Took approximately 8 hours per fold
3. Results computed after each fold

| Set | Count of citations |
|---|---|
| Training | 47853 |
| Validation | 15947 |
| Test | 15970 |

Input data statistics

Classfication Performance for 1821060:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| author | 1.00 | 1.00 | 1.00 | 132368 |
| citation-number | 0.99 | 1.00 | 0.99 | 8880 |
| collection-editor | 0.41 | 0.97 | 0.58 | 411 |
| collection-title | 0.90 | 0.99 | 0.94 | 45070 |
| container-author | 0.71 | 0.02 | 0.03 | 630 |
| container-title | 0.98 | 0.94 | 0.96 | 105156 |
| doi | 1.00 | 0.99 | 1.00 | 10613 |
| editor | 0.72 | 0.67 | 0.69 | 1875 |
| extra | 1.00 | 1.00 | 1.00 | 104612 |
| interviewer | 0.65 | 0.56 | 0.60 | 290 |
| issue | 1.00 | 0.98 | 0.99 | 8793 |
| issued | 0.99 | 0.97 | 0.98 | 5886 |
| month | 0.99 | 1.00 | 0.99 | 2407 |
| number | 0.99 | 0.99 | 0.99 | 797 |
| page | 0.97 | 0.99 | 0.98 | 8609 |
| publisher | 0.97 | 1.00 | 0.98 | 21442 |
| punctuation | 0.97 | 0.98 | 0.97 | 1359 |
| title | 1.00 | 0.99 | 0.99 | 183507 |
| url | 1.00 | 0.99 | 0.99 | 5313 |
| volume | 0.99 | 0.98 | 0.99 | 13673 |
| year | 0.97 | 1.00 | 0.98 | 12267 |
| micro avg | 0.98 | 0.98 | 0.98 | 673958 |
| macro avg | 0.91 | 0.90 | 0.89 | 673958 |
| weighted avg | 0.98 | 0.98 | 0.98 | 673958 |

# Evaluation

- Gold Tag Discussion
  - Multi-correct labels?
  - Gold set not so gold
  - Different but equal tags
- Formatting For Evaluation
  - Tokenization/parsing, labels
  - Binary, strings
  - Label match ups

Table 4.8: False Negative Cases Identified

| Gold Tags | False Negative Prediction | Check Implemented |
|---|---|---|
| authors | author | YES |
| booktitle/journal | container-title | YES |
| address | location | NO |
| month/year | date | NO |

# Evaluation

- Input: Per Gold Tag, Cumulative
  - Per tag affected scores when using sklearn

- 21 categories, 8 gold
  - F1, Precision, and Recall calculated for each

```
4    publisher   publisher
5    publisher   publisher
6    publisher   publisher
7    publisher   extra
8    publisher   collection-title
9    publisher   collection-title
10   publisher   container-title
11   publisher   publisher
12   publisher   publisher
13   publisher   publisher
14   publisher   citation-number
15   publisher   container-title
16   publisher   container-title
```

```
map_dict={'authors': 0, 'booktitle': 1,  'editor': 2, 'journal': 3, 'publisher': 4, 'title': 5, 'volume': 6, 'year': 7,
          'citation-number': 8, 'author': 9, 'container-title': 10, 'punctuation': 11, 'NA': 12, 'collection-editor': 13,
          'container-author': 14, 'issue': 15, 'extra': 16, 'collection-title': 17, 'page': 18, 'issued': 19, 'doi': 20}
```

# Discussion

1. Tags like “title” and “booktitle” perform decently
2. “Author” and “editor” are performing poorly
3. Publisher is also classified as “collection-title” and “container-title”
4. “Year” performs poorly

| Gold Tags | F1 | P | R |
|---|---|---|---|
| title | 0.8454 | 0.7844 | 0.9167 |
| editor | 0.2143 | 0.1351 | 0.5172 |
| booktitle | 0.7560 | 0.8462 | 0.6832 |
| journal | 0.6948 | 0.8425 | 0.5912 |
| authors | 0.6803 | 0.9405 | 0.5329 |
| volume | 0.2953 | 0.3143 | 0.2785 |
| publisher | 0.3107 | 0.4444 | 0.2388 |
| year | 0.0635 | 0.1667 | 0.0392 |

Evaluation against gold standard

# Examples

<author>Abate, Alessandro, et al.</author>. <title>“Box Invariance for Biologically-Inspired Dynamical Systems.”</title> <container-title>**2007 46th IEEE Conference on Decision and Control**</container-title>, <publisher>**IEEE**</publisher>, <issued>**2007**</issued>, <DOI>doi:10.1109/cdc.2007.4434569</DOI>.

<author>A.M. Lätti and H. Koskela and J. Pekkanen</author>. <title>“Prolongation of Recent-Onset Cough: A Prospective Follow-Up Study Among Finnish Adult Employee Population.”</title> <container-title>**C37. SYMPTOMS, PLEURAL DISEASE, BEHAVIORAL SCIENCE, AND OTHER TOPICS**</container-title>, <publisher>American Thoracic Society</publisher>, <issued>**2019**</issued>, <DOI>doi:10.1164/ajrccm-conference.2019.199.1_meetingabstracts.a4699</DOI>.

# Figure Extraction

# DeepFigures

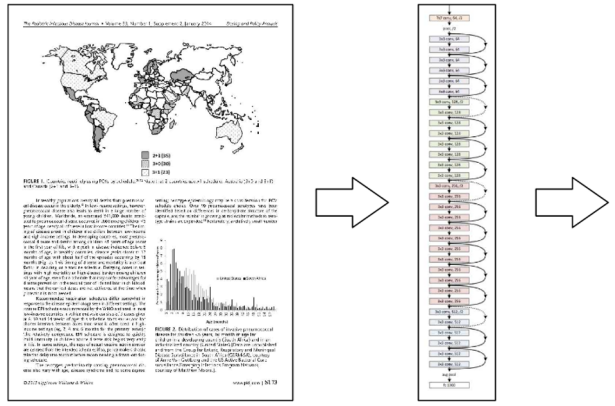
Novel method to extract non-textual components (figures and tables) from PDF

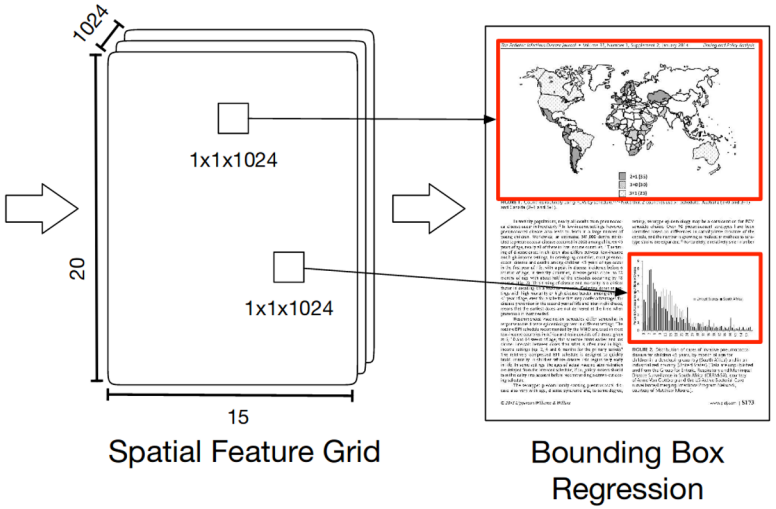
Two datasets were used for training—arXiv and PubMed

For arXiv, authors modified the original LaTeX source code of the PDFs such that it rendered bounding boxes around the figures and captions

For PubMed, the figures were already present in parsed format

These induced labels were then used to train a deep learning model to predict the coordinates of the bounding boxes around figures




Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting Scientific Figures with Distantly Supervised Neural Networks. CoRRabs/1804.02445 (2018). arXiv:1804.02445 Retrieved October 9, 2019 from http://arxiv.org/abs/1804.02445

# Figure Extraction: Born Digital vs. Scanned

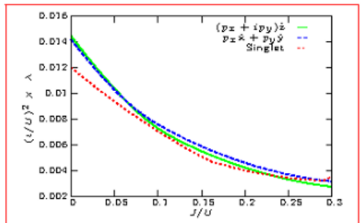


FIG. 5: RG eigenvalue λ for the chiral state, the most favoured helical state and the singlet state for the parameters given in the main text.

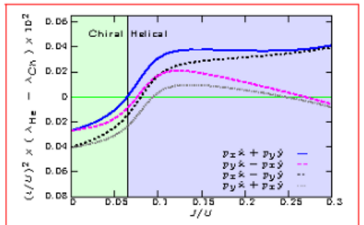


FIG. 6: Splitting of the RG eigenvalue λ between the four different helical states and the chiral state for the parameters given in the main text.



6

magnitude on $\gamma$ while the helical state at larger $J/U$ has a larger gap amplitude on $\alpha$ and $\beta$.

In summary, the amplitude of $\eta$ does not modify qualitatively our findings of a favoured chiral state with a (slightly) dominant $\gamma$ for small $J/U$ and a favoured helical state with (slightly) dominant $\alpha$ and $\beta$ for larger $J/U$.

$J/U$. As figured by the arrows in Fig. 7, increasing $\eta$ does mostly two things: it increases the amplitude of the splitting (be it positive or negative) in the two aforementionned regions and slightly increases the value of $J/U$ at which the cross over happens. It also makes the cross over smoother. In the limit $\eta \to 0$, the splitting would go to zero, from below in the former region and from above in the latter.

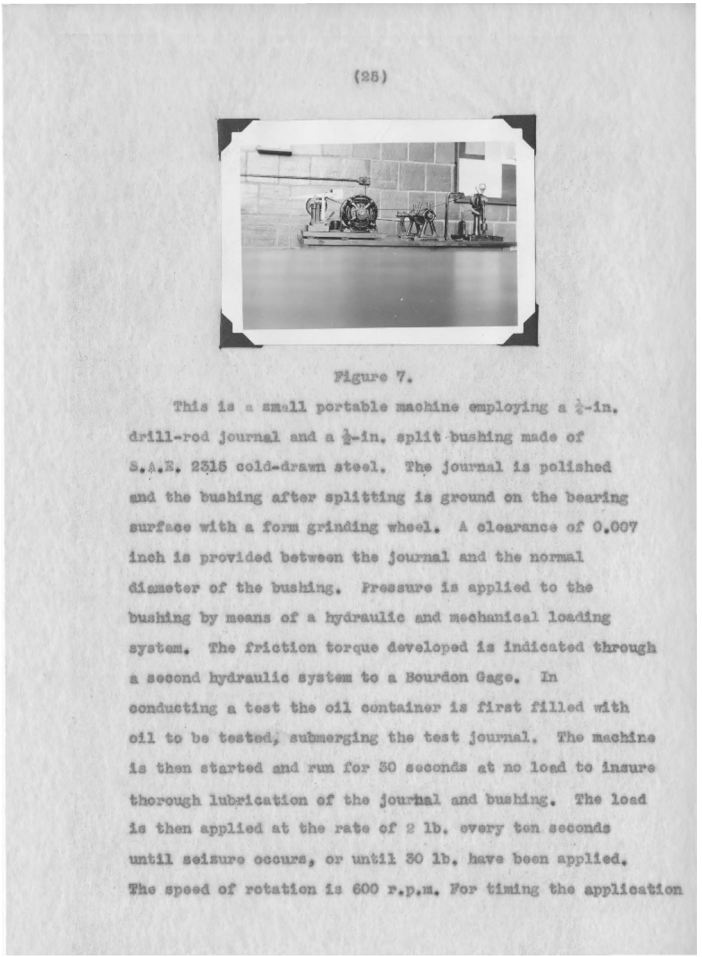
The ratio of the maxima of the gap amplitudes over the different bands $R = \frac{\max|\Delta_{\alpha,\beta}|}{\max|\Delta_\gamma|}$ for different SOC parameters $\eta$ is shown in Fig. 8. Regardless of the value of $\eta$, the chiral state favoured at small $J/U$ has a larger gap



FIG. 7: Splitting of the RG eigenvalues λ between the chiral and the helical state for different SOC parameters $\eta$. All the other parameters are given in the main text.



FIG. 8: Ratio of the maxima of the gap amplitudes over the different bands $R = \frac{\max|\Delta_{\alpha,\beta}|}{\max|\Delta_\gamma|}$ for different SOC parameters $\eta$. All the other parameters are given in the main text. At each value of $J/U$, only the curve for the most favoured state (chiral or helical) is shown. "Ch" stands for chiral and "He" stands for helical.



(25)

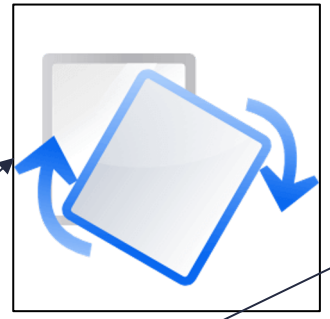
Figure 7.

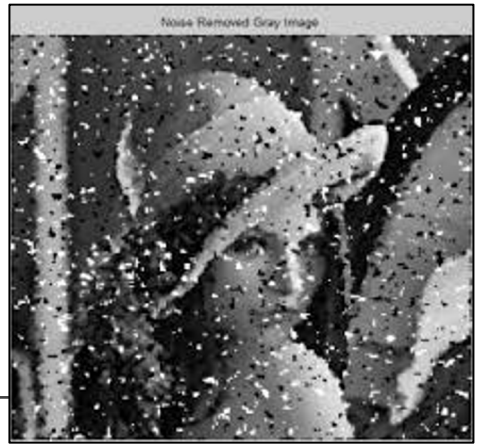
This is a small portable machine employing a ½-in. drill-rod journal and a ½-in. split bushing made of S.A.E. 2315 cold-drawn steel. The journal is polished and the bushing after splitting is ground on the bearing surface with a form grinding wheel. A clearance of 0.007 inch is provided between the journal and the normal diameter of the bushing. Pressure is applied to the bushing by means of a hydraulic and mechanical loading system. The friction torque developed is indicated through a second hydraulic system to a Bourdon Gage. In conducting a test the oil container is first filled with oil to be tested, submerging the test journal. The machine is then started and run for 30 seconds at no load to insure thorough lubrication of the journal and bushing. The load is then applied at the rate of 2 lb. every ten seconds until seizure occurs, or until 30 lb. have been applied. The speed of rotation is 600 r.p.m. For timing the application

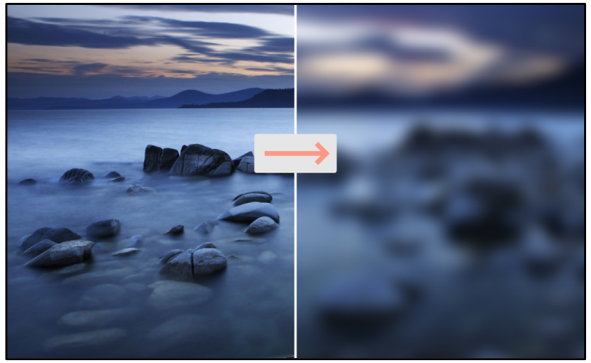
Image 1 src: https://arxiv.org/pdf/1804.02445.pdf
Image 2 src: https://vtechworks.lib.vt.edu/handle/10919/56159
Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting Scientific Figures with Distantly Supervised Neural Networks. CoRRabs/1804.02445 (2018). arXiv:1804.02445 Retrieved October 9, 2019 from http://arxiv.org/abs/1804.02445

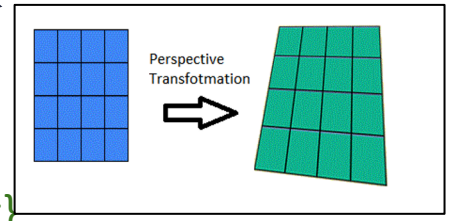
# Methods

1. Image-based data augmentations
   a. Random affine rotation (limited to +/- 5 degrees)
   b. Additive Gaussian noise
   c. Salt-and-pepper noise
   d. Gaussian blur
   e. Linear contrast
   f. Perspective transform
2. Latex-based data augmentations
   a. \documentclass[sigconf]{acmart}
      \documentclass[sigconf,12pt]{acmart}
   b. Following code added at the beginning:
      \renewcommand\ttdefault{cmvtt}
      \renewcommand{\familydefault}{\ttdefault}
      \linespread{1.5}







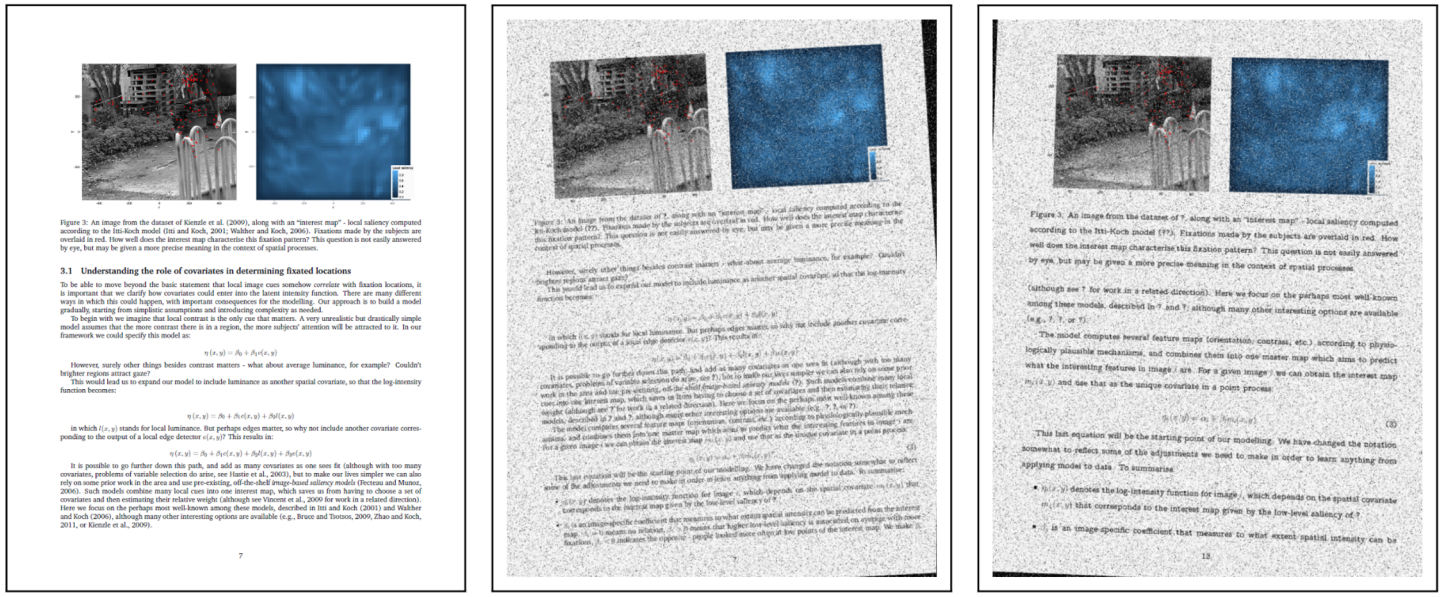


Low Contrast Image


High Contrast Image

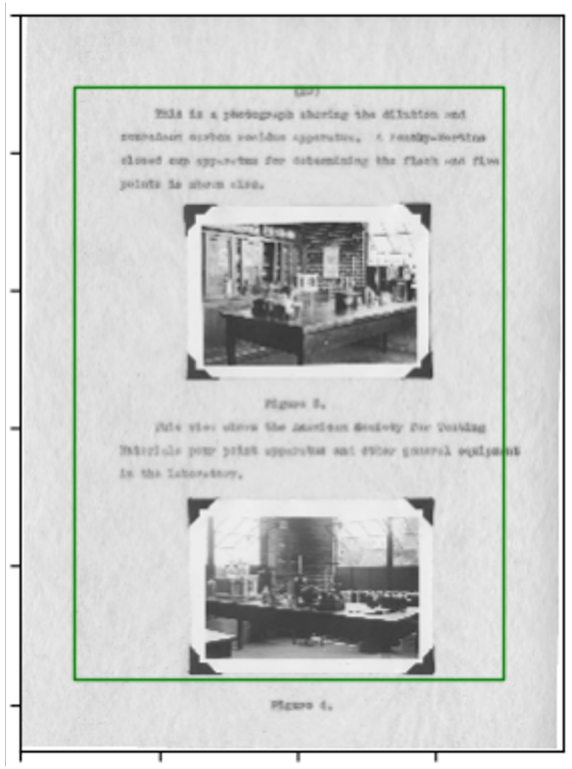
# Experiments

Train the DeepFigures model on our augmented data set

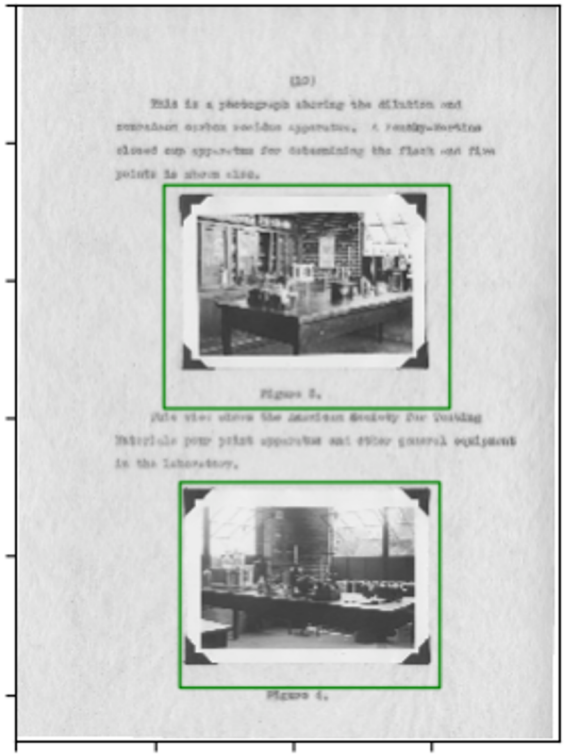


Simon Barthelmé, Hans Trukenbrod, Ralf Engbert, and Felix Wichmann. 2012. Modelling fixation locations using spatial point processes. arXiv:stat.AP/1207.2370

# Results and Discussion



Original model

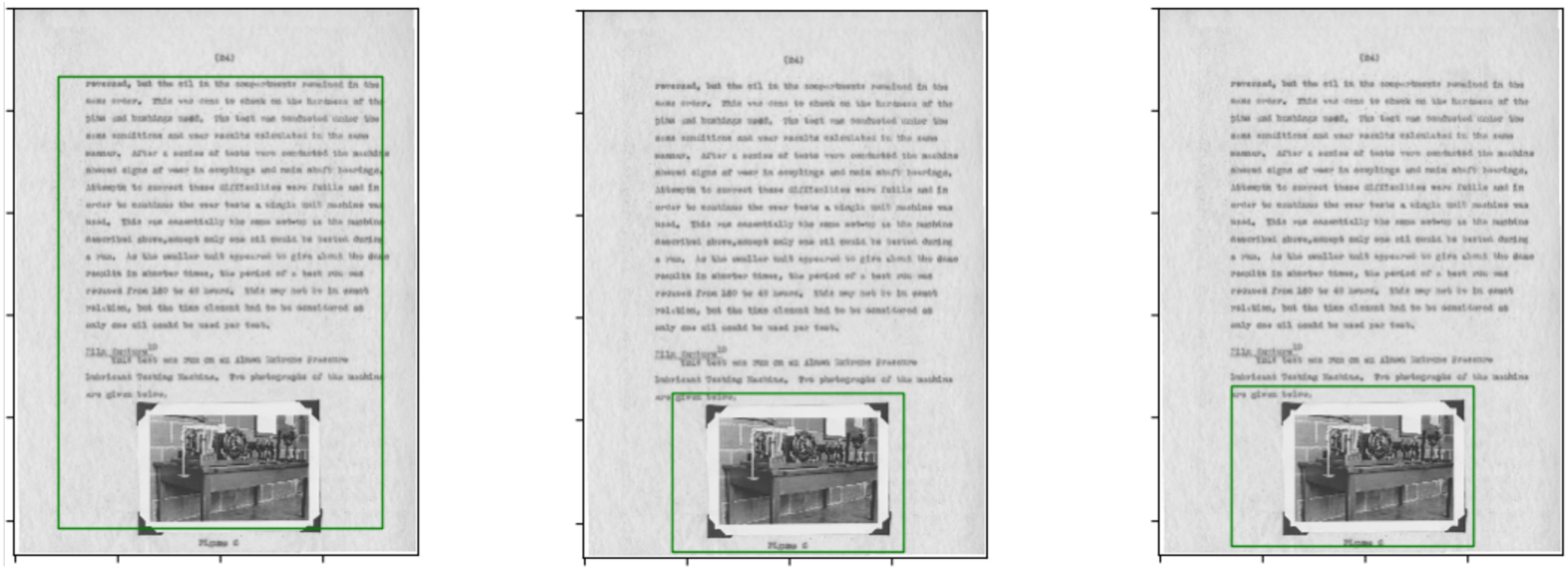


(Ours) Model trained on image-based transformations

(Ours) Model trained on all transformations

ETD source: Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159

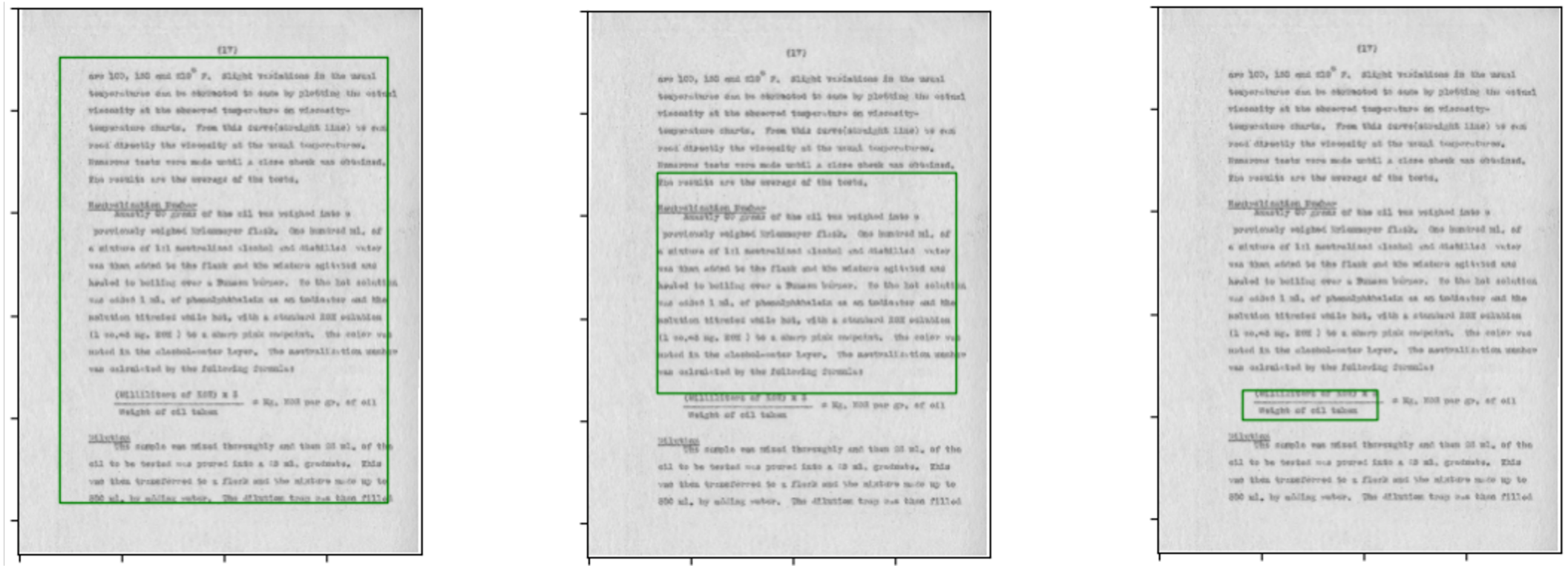
# Results and Discussion



Original model (Ours) Model trained on image-based transformations (Ours) Model trained on all transformations

ETD source: Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159

# Results and Discussion



Original model (Ours) Model trained on image-based transformations (Ours) Model trained on all transformations

ETD source: Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159

# Evaluation

Evaluated the model on a single ETD from VTechWorks.
Number of actual images: 26 (ground truth)

| | Original model (DeepFigures) | Model trained on image-based transformations | Model trained on all transformations |
|---|---|---|---|
| True Positives (correct predictions) (Higher is better) | 0 | 7 | 10 |
| False positives (incorrect predictions) (Lower is better) | 29 | 16 | 15 |
| False negatives (missed predictions) (Lower is better) | 26 | 21 | 16 |

ETD source: Walter Douglas Chiles. 1935. Effect of service on automobile crankcase oils. Ph.D. Dissertation. Virginia Agricultural and Mechanical College and Polytechnic Institute. http://hdl.handle.net/10919/56159

# Future Work

# Future Work

**Classification**

- Use Word Embeddings like GloVe, BERT along with Deep Learning models
- Extend the work to include full-text data
- Extend these methods to classify ETD chapters

**Parsing**

- Pre-processing of data to be done carefully
- Train on more data and types of documents
- How golden are the gold tags?

**Figure Extraction**

- Train on more augmented data.
- Try more augmentations.
- Try more architectures (YOLOv3, SSD, etc)

# Thank you



This research was done under the supervision of Dr. Edward A. Fox

as part of the course

CS6604: Digital Libraries at Virginia Tech, Fall 2019.

https://github.com/waingram/CS6604-ETD/