

ETDseer

Final Term Term Project Presentation

Yufeng Ma, Tingting Jiang, Chandani Shrestha

CS 6604 - Digital Libraries

Virginia Tech, Blacksburg, VA, 24061

Professor Edward A. Fox

April 27, 2017



Overview

- **Problem Statements/Motivation**
- **Related Works**
 - NDLTD
 - CiteSeer
- **Stakeholder Overview**
 - **Architecture**
 - **Scenarios and Services**
 - Workflow example
- **Structured Data Extraction**
 - ETD Segmentation
 - Table & Figure Extraction
 - Reference Extraction
- **Text Summarization**
- **Network Visualization**
- **Working On...**



Background

Problem Statements

- ETDs as international resource- extensive potential
- Largely Untapped
- Limitation in existing related tools
 - Document length, Document accessibility (Full Text)
 - Summarization, Visualization
- Challenges working with ETDs
- Need for enhanced usability

Motivation

- Rich knowledge base - Single platform
- Accessible to broader group of users
- Network of institutional repositories



Related Works

- NDLTD
- VTechworks
- CiteSeer
- ContentMine

VTechWorks

NDLTD

NDLTD Search this site:

News Videos Community Thesis Resources Global ETD Search

Submit Nominations for 2017 ETD Awards by May 31st, 2017

NDLTD is pleased to announce the 2017 ETD Awards program. We invite all NDLTD members to nominate individuals they feel deserve the recognition!

Deadline for Nominations: May 31st, 2017

The NDLTD's ETD Awards recognize and support innovative theses and dissertations and leadership within the ETD community. These awards are presented each year at the annual ETD Symposium.

The awards include several categories of appreciation:

- The Innovative ETD Award supports student efforts to transform the genre of the dissertation through the use of innovative research data management techniques and software to create multimedia ETDs.
- The ETD Leadership Award recognizes individuals whose leadership and vision has helped raise awareness of the benefits of open access ETDs and whose efforts have improved graduate education and research through the use of technology.

The awards will be presented at ETD 2017 "Exploring Global Connections" the 20th International Symposium on Electronic Theses and Dissertations, Washington, DC August 7-9, 2017.

Award Category Information, Procedures and Requirements:

Innovative ETD Award

The intent of this award is to support current graduate students in the creation of innovative ETDs. We anticipate students will use the award to assist in the purchase of specialized software, computer equipment or other technical support needed in the production or publication of their ETD. Nominations for this award should indicate how the award will support the application and integration of renderings, photos, data sets, software code or other multimedia objects in the student's ETD.

- A prospective thesis or dissertation may be nominated by a representative (e.g. administrator, faculty member or librarian) of an institution that is a member of the NDLTD.
- Each NDLTD member may submit up to 2 nominees. In addition, a student from a member organization may submit a self-nomination.
- The nomination should include their approved thesis proposal, along with a brief statement about how the work will use technology to enhance the presentation of the research in the forthcoming ETD.
- The resulting ETD must be distributed online as open access upon publication.
- Students must provide an ORCID ID upon acceptance of the award.
- There may be up to three winners in this category.
- Each winner will receive \$1,000.
- Winners are required to record a video acceptance message (approximately 2 - 3 minutes) that may be presented at the annual symposium's awards ceremony and distributed online, or provide an acceptance message in writing (500 words or less).

Nominations may be made through the following online form: <http://apo.oi/forms/ARF5u5c2>

ETD Leadership Award

Log in

Virginia Tech Home / ETDs: Virginia Tech Electronic Theses and Dissertations

ETDs: Virginia Tech Electronic Theses and Dissertations

BROWSE BY

By Issue Date Authors Titles Subjects

Search within this community and its collections:

Go

etdsvt

Virginia Tech has been a worldwide leader in electronic theses and dissertation initiatives for more than 20 years. On January 1, 1997, Virginia Tech was the first university to require electronic submission of theses and dissertations (ETDs). Ever since then, Virginia Tech graduate students have been able to prepare, submit, review, and publish their theses and dissertations online and to append digital media such as images, data, audio, and video.

University Libraries staff are also currently digitizing thousands of pre-1997 theses and dissertations and loading them into VTechWorks. As of February 2017, most of these theses and dissertations are fully available to the public, but we will in general honor requests by the item's author to restrict access to Virginia Tech only. See our process for Requesting that Material be Amended or Removed.

Materials that are restricted to Virginia Tech only may be requested via your own university or public library's Interlibrary Loan program or through the VTechWorks request form that appears when you try to access the item. You might also be able to obtain a copy of the work through ProQuest's database of theses and dissertations, if you are on a Virginia Tech campus but are unable to find the pre-1997 thesis or dissertation you are seeking in VTechWorks, you may also be able to order a physical copy from library storage. Please check the library catalog at <http://www.lib.vt.edu/> for physical copies.

The guidelines that apply to Virginia Tech's graduate students as ETD authors can be found at <http://guides.lib.vt.edu/ETDguide>.

Collections in this community

Award-winning Theses and Dissertations [5]

Doctoral Dissertations [12464]

Masters Theses [17500]

VT ETD Resources [11]

Documentation about creating and formatting VT ETDs.

Recent Submissions

Methanocaldicoccus jannaschii and the Recycling of S-adenosyl-L-methionine
 Miller, Daniele Virginia (Virginia Tech, 2017-04-26)
 S-Adenosyl-L-methionine (SAM) is an essential metabolite for all domains of life. SAM-dependent reactions result in three major metabolites: S-

CiteSeerX

Documents Authors Tables Donate MetaCart Sign up Log in

Include Citations Advanced Search

Most Cited: Documents, Citations, Authors

Powered by: **Solr**

[About CiteSeerX](#)
[Submit and Index Documents](#)
[Privacy Policy](#)
[Help](#)
[Data](#)
[Source](#)
[Contact Us](#)

Developed at and hosted by [The College of Information Sciences and Technology](#)

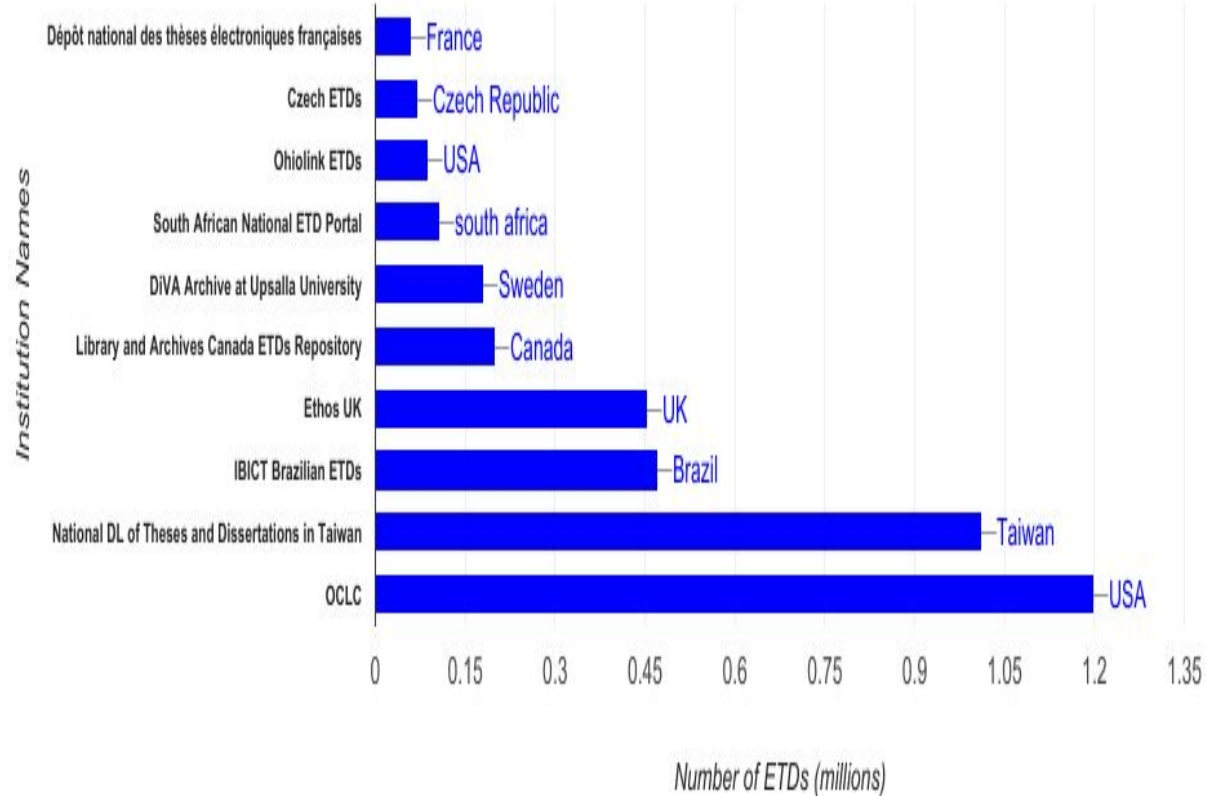
© 2007-2016 The Pennsylvania State University



NDLTD

- ETDs Statistics
- Functional Limitations
 - Categorization Missing
 - No Full Text Access

ETDs Distribution over NDLTD





CiteSeerX

- Data: Make use of Metadata
- Technologies: SeerSuite tools
 - Automatic citation indexing
 - Automatic metadata extraction
 - Reference linking
 - Author disambiguation
 - Related documents
 - Full-text indexing

Documents Authors Tables [Donate](#) [MetaCart](#) [Sign up](#) [Log in](#)







CiteSeer^x 10M

author:(Edward A. Fox) AND affil:(Virginia Tech) AN  
 Include Citations [Advanced Search](#)

Results 1 - 10 of 15,703 [Next 10 →](#)

[Tools](#)

Sorted by:
[Citation Count](#) ↓

Try your query at:
  
  

[Modern Information Retrieval](#)
by Ricardo Baeza-Yates, Berthier Ribeiro-Neto , 1999
"... Information retrieval (IR) has changed considerably in the last years with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become quite out-of-date which has led to the i ..."
Abstract - Cited by 3159 (30 self) - [Add to MetaCart](#)

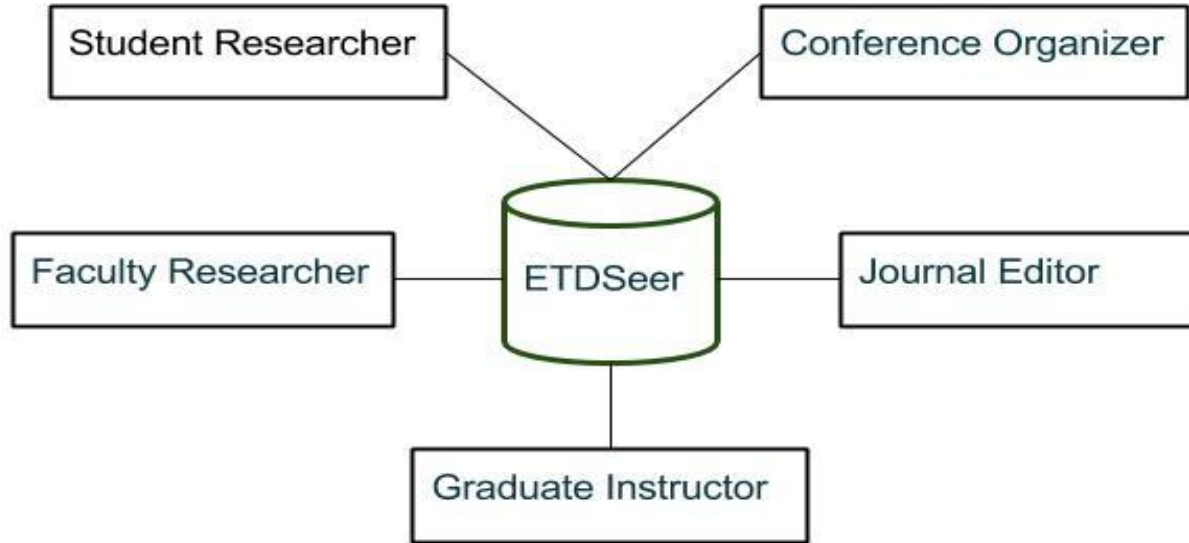
[Planning Algorithms](#)
by Steven M LaValle , 2004
"... This book presents a unified treatment of many different kinds of planning algorithms. The subject lies at the crossroads between robotics, control theory, artificial intelligence, algorithms, and computer graphics. The particular subjects covered include motion planning, discrete planning, planning ..."
Abstract - Cited by 1111 (53 self) - [Add to MetaCart](#)

[Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev](#)
by R I Amann, W Ludwig, K H Schleifer, Rudolf I. Amann, Wolfgang Ludwig, Karl-heinz Schleifer , 1995
"... cultivation of individual microbial cells without Phylogenetic identification and in situ detection ..."
Abstract - Cited by 1070 (29 self) - [Add to MetaCart](#)

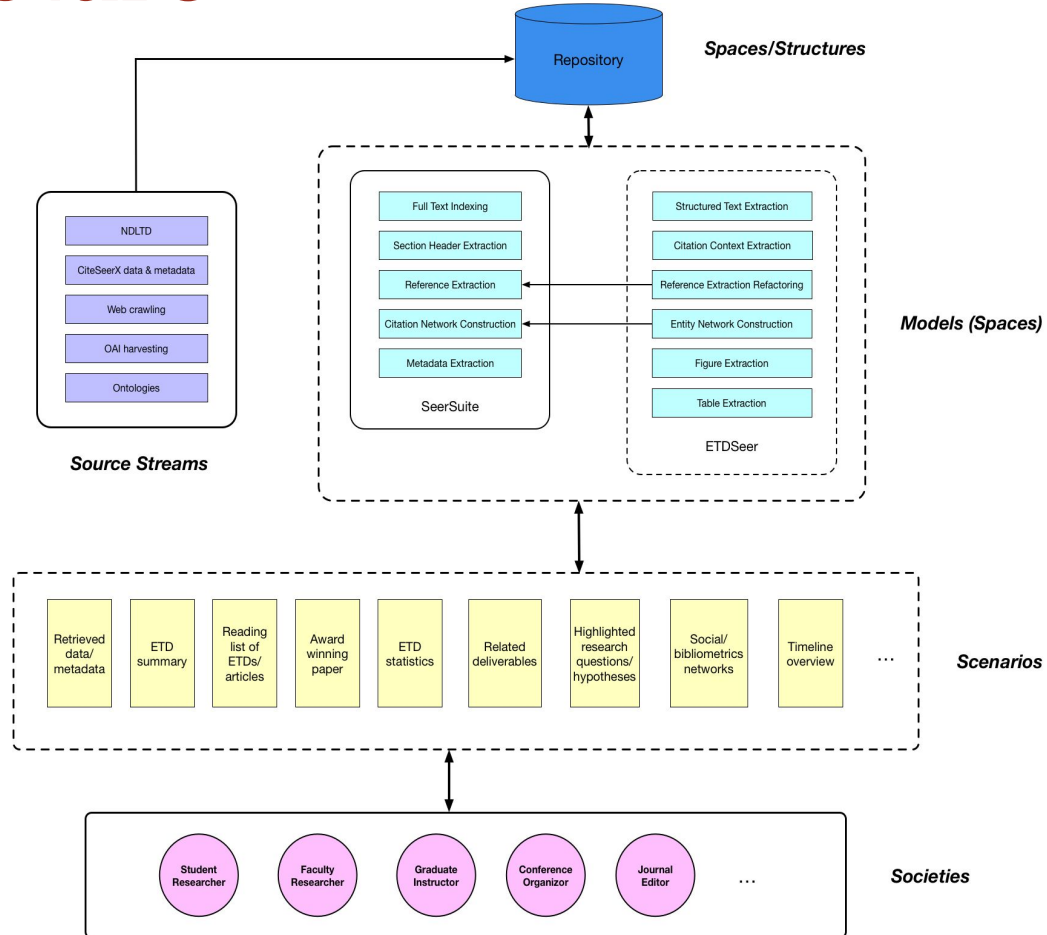
[SPINS: Security Protocols for Sensor Networks." Wireless Networks 8](#)
by Adrian Perrig, Robert Szewczyk, Victor Wen, David Culler, J. D. Tygar
"... As sensor networks edge closer towards wide-spread deployment, security issues become a central concern. So far, the main research focus has been on making sensor networks feasible and useful, and less emphasis was placed on security. We design a suite of security building blocks that are optimized ..."
Abstract - Cited by 1052 (32 self) - [Add to MetaCart](#)



Stakeholders Overview



Architecture

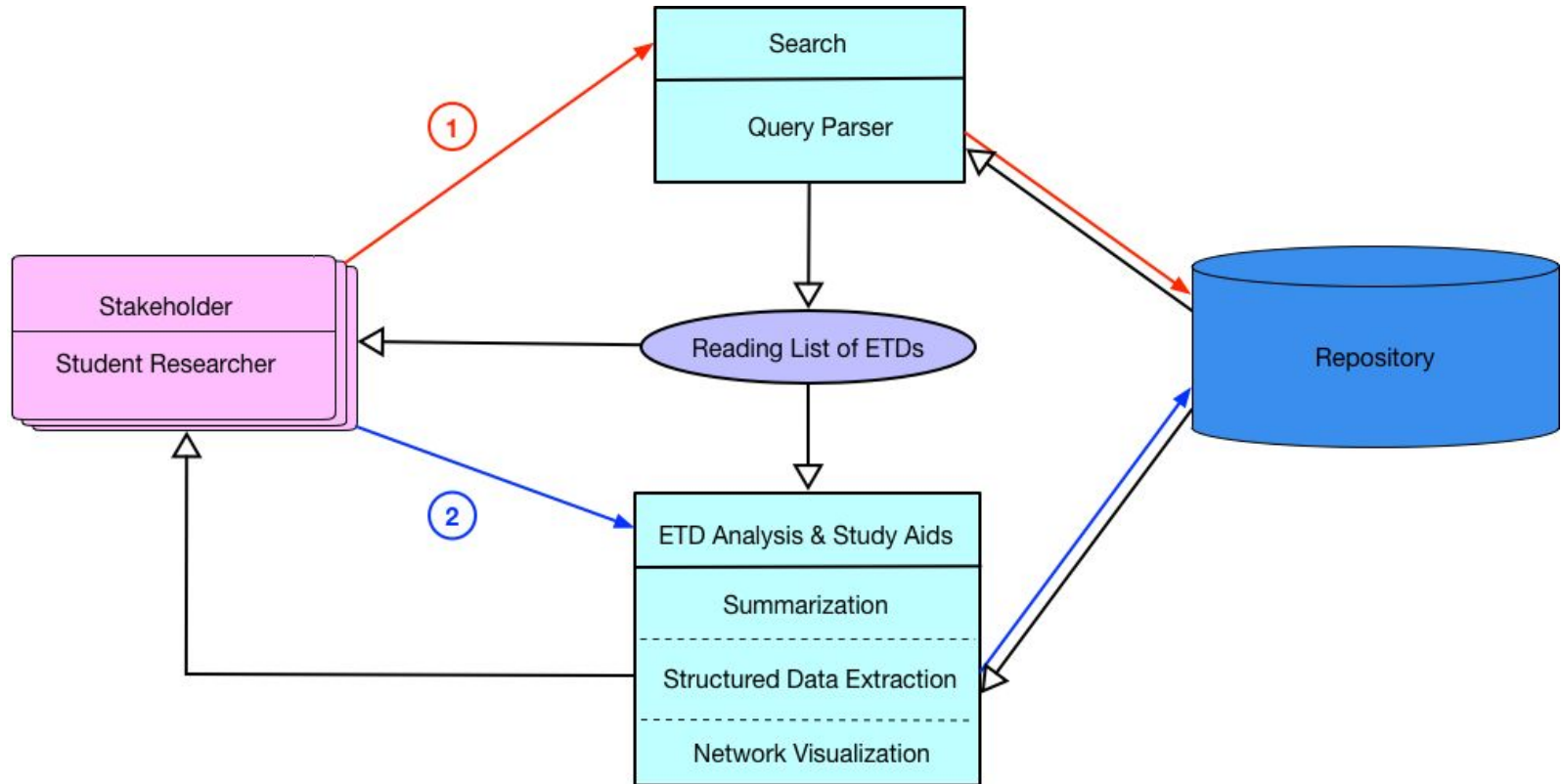


Scenario 1 - Student Researcher

Requirements	Key techniques	Expected Outputs
Metadata-based search	SeerSuite	<ul style="list-style-type: none">• Specific ETDs within a date range• Specific ETDs with an advisor name
Research interests discovery	Structured text extraction	<ul style="list-style-type: none">• Desired ETDs with quality scores• Research questions/hypotheses highlighted
Reference extraction	Structured text extraction	<ul style="list-style-type: none">• Related ETDs/books/articles/papers• Tabular/Canonical representations• Downloadable package of related work• Lists of journals/conferences
Linking of problems with methods	Text extraction	<ul style="list-style-type: none">• Different methods for a problem• A site with detailed resources• An award winning paper (outline/draft)
ETD analysis and study aids	Deep learning	<ul style="list-style-type: none">• ETD content summarizations• Figures, tables and equations• Key sections and list of related problems• Visualizations (social/bibliometrics networks)• Timeline overview of evolutionary work



An Example Workflow



Other Scenarios

Stakeholder	Requirements	Expected Outputs
Faculty Researcher	Research problem exploration aid	<ul style="list-style-type: none">● Synthesis of related ETDs● Proposed approaches/solutions● Future works summarization
Graduate Instructor	Graduate course syllabus formulation	<ul style="list-style-type: none">● Draft with a hierarchical topical outline● Link to each topical entry with a reading list
	Advanced topic, lecture preparation	<ul style="list-style-type: none">● Slides cover research questions/problems● Synthesis of provided potential solutions



Other Scenarios

Stakeholder	Requirements	Expected Outputs
Conference Organizer	TPC member identification	<ul style="list-style-type: none">• List of advisor research faculty names• Ranking table of advisors
	Potential participants identification	<ul style="list-style-type: none">• Subgraph of ETD-derived citation graph• CSV file of author names, contact info.
Journal Editor	Peer-reviewer identification	Research interest-based reviewer list
	Content originality check	<ul style="list-style-type: none">• Previous publications of the authors• Estimated percentage of the new content/work



Structured Data Extraction: ETD Segmentation

- **Heuristics-based strategy**
 - Start with 'Chapter' or 'CHAPTER'
 - Font size and style
- **Deep learning approaches**
 - Treat every two pages as one image
 - Manually label each image - chapter breaking point or not
 - Build a CNN model for classification

Chapter 4

Recursive Composition Functions

The previous chapter introduced standard RNNs and introduced the main objective functions. This chapter investigates more powerful RNN architectures that move beyond having the same, standard neural network for composing parent vectors as in the previous chapter. The main objective functions, I explored are

1. Syntactically untied RNNs: The composition matrix that is used to compute the parent vector depends on the syntactic category of the children.

Figure 8.2 Average number of connections per neuron in neural and machine learning models. Part of the success of machine learning models despite their low number of neurons may be due to the comparatively high number of connections between neurons in machine learning models. In fact, machine learning models are not far from human levels of connectivity. This suggests that techniques that increase the total number of features in a model while reducing dimensionality (e.g. feature selection) may be very effective. This explains the success of models that employ sparse connectivity and pooling, such as convolutional networks, especially recurrent networks (see chapter 9). Estimate of the average number of connections per neuron obtained by dividing the number of neurons by the number of synapses listed at http://en.wikipedia.org/wiki/List_of_brains_by_weight_of_neurons, the number of neurons, "RNN" refers to (Hinton et al., 2010), "Adaptive" refers to (Sutton et al., 2010). Images, with the exception of the photo of my son, Philipp, are not my own.

35

4 Prologue to First Article

4.1 Article Details

Scaling up Spike-and-Slab Models for Unsupervised Feature Learning. Ian J. Goodfellow, Aaron Courville, and Yoshua Bengio. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8), 1902-1914.

Personal Contributions. The idea that a structural variational inference in a sparse coding model could provide an effective means of feature extraction was my own idea. Aaron Courville suggested using spike-and-slab sparse coding as the base model; my original idea was to use binary sparse coding. Aaron Courville also suggested one of the two inference algorithms presented in the paper, the method based on conjugate gradient descent. Aaron Courville and I developed the equations necessary for inference and learning jointly. The partially directed deep Boltzmann machine was my own idea. I implemented all of the necessary software and performed all of the experiments. I wrote the majority of the paper, with significant contributions to the writing from both Aaron Courville and Yoshua Bengio. I produced all of the figures.



Structured Data Extraction: Table & Figure Extraction

- Extend work from TableSeer
 - Table box detection
 - Table metadata extraction
 - Deal with styles of more disciplines
- Figure Extraction
 - Sagnik Ray Choudhury's work
 - CNN based approach - Mask R-CNN
 - AMT for figure labeling

- * K. He, et al., Mask R-CNN. *arXiv*, 2017.
- * Y. Liu, et al., TableSeer, *JCDL* 2007.
- * S. R. Choudhury, et al., An Architecture for information extraction from figures in DLs, *WWW*, 2015.

Table 9.1 Test set misclassification rates for the best methods on the permutation invariant MNIST dataset. Only methods that are regularized by modeling the input distribution output from the maxout MLP.

METHOD	TEST ERROR
RECTIFIER MLP + DROPOUT (SERVATANA, 2013)	1.05%
DHM (SALAKUTDINOV AND HINTON, 2009)	0.95%
Maxout MLP + dropout	0.94%
MP-DDM (GOODFELLOW ET AL., 2013)	0.88%
DEEP CONVEX NETWORK (YU AND DING, 2011)	0.83%
MANIFOLD TANGENT CLASSIFIER (RIZAL ET AL., 2011)	0.81%
DHM + DROPOUT (HINTON ET AL., 2012)	0.79%

9.5.1 MNIST

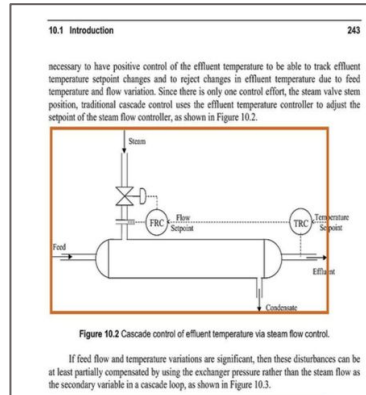
The MNIST (LeCun et al., 1998) dataset consists of 28×28 pixel greyscale images of handwritten digits 0-9, with 60,000 training and 10,000 test examples. For the permutation invariant version of the MNIST task, only methods unaware of the 2D structure of the data are permitted. For this task, we trained a model consisting of two densely connected maxout layers followed by a softmax layer. We regularized the model with dropout and by imposing a constraint on the norm of each weight vector, as in (Srebro and Shraibman, 2005). Apart from the maxout units, this is the same architecture used by Hinton et al. (2012). We selected the hyperparameters by minimizing the error on a validation set consisting of the last 10,000 training examples. To make use of the full training set, we recorded the

```

<table-metadata>
<property>
<name>Paper Title</name>
<value>Dissolution of albite glass and crystal</value>
</property>


<property>
<name>Table Caption</name>
<value>Table 2. Comparison of crystalline and amorphous albite dissolution rates</value>
</property>

<property>
<name>Table Column Head</name>
<value>Type of experiment Initial pH Final pH Temperature</value>
<description>.....</description>
</property>
.....
</table-metadata>
    
```



Structured Data Extraction: Reference Extraction

- References that appear at the end
 - SeerSuite
- References that appear anywhere like footnotes
 - Deep learning for learning reference features
 - Classifier will be trained
- Represented in canonical format like BibTeX



Bibliography

Abharif, O. and J. Pineau (2013). End-to-end text recognition with hybrid HMM maxout models. Technical report, arXiv:1310.1811.

Arnold, L. and Y. Ollivier (2012, December). Layer-wise learning of deep generative models. Technical report, arXiv:1212.1524.

Bastien, F., P. Lamblin, R. Pascanu, J. Bergeron, I. J. Goodfellow, A. Bergeron, N. Brechard, and Y. Bengio (2012). Thesno: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning* 2(1), 1–127. Also published as a book. Now Publishers, 2009.

Bengio, Y., P. Lamblin, D. Popovici, and H. Larochelle (2007). Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pp. 153–160. MIT Press.

Bengio, Y., Y. LeCun, C. Nohl, and C. Burges (1995). Leroc: A NN/HMM hybrid for on-line handwriting recognition. *Neural Computation* 7(6), 1289–1303.

Bengio, Y., E. Thibodeau-Lafeur, C. Alain, and J. Yosinski (2014). Deep generative stochastic networks trainable by backprop. Technical Report arXiv:1306.1091.

Donald E. Knuth's¹ Principal Principle is one of my favourite principles. One of my favourite books is *T_EXbook*. Everybody should be rational.² Knuth said a lot of things. For instance, he said that everybody should be rational,³ and he said that everybody should drive on the right side of the road.⁴ Arnold van Gennep said that everybody should drive on the left, but otherwise van Gennep's⁵ work agrees with Knuth.

¹Donald E. Knuth. *Computers & Typesetting*. Vol. A: *The T_EXbook*. Reading, Mass.: Addison-Wesley, 1984.

²Knuth, *T_EXbook*; Arnold van Gennep. *Les rites de passage*. Paris: Nourry, 1909.

³Knuth, *T_EXbook*, p. 9.

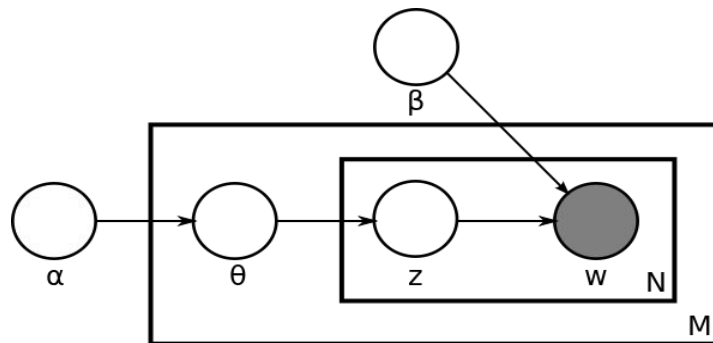
⁴Donald E. Knuth. *Computers & Typesetting*. Vol. C: *The METAFONTbook*. Reading, Mass.: Addison-Wesley, 1986, pp. 10–15.

⁵Van Gennep, *Rites de passage*, p. 4.

Text Summarization

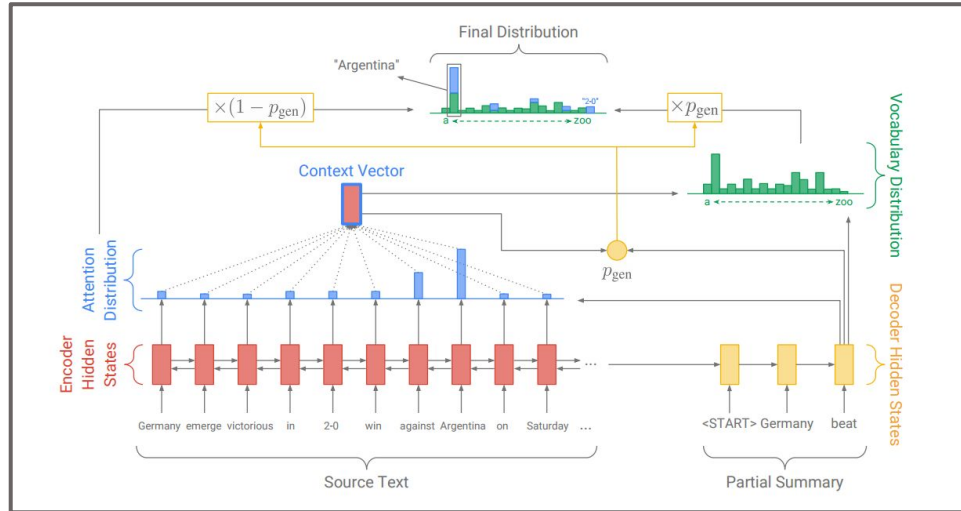
- **Topic modeling based approach**

- Extracted keyword or phrase
- Probabilistic graphical model



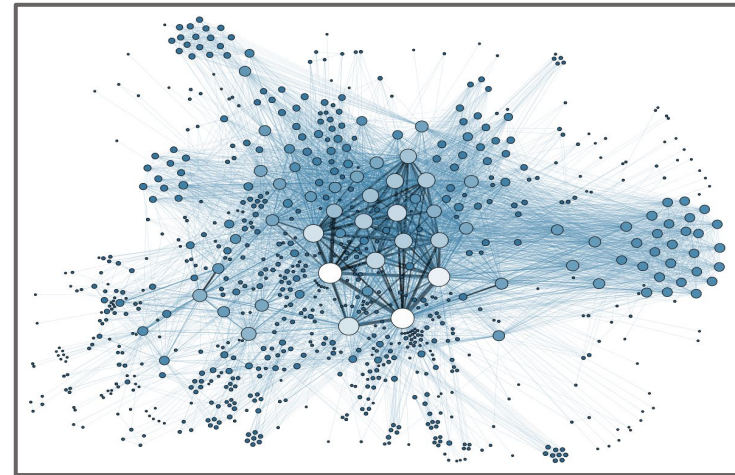
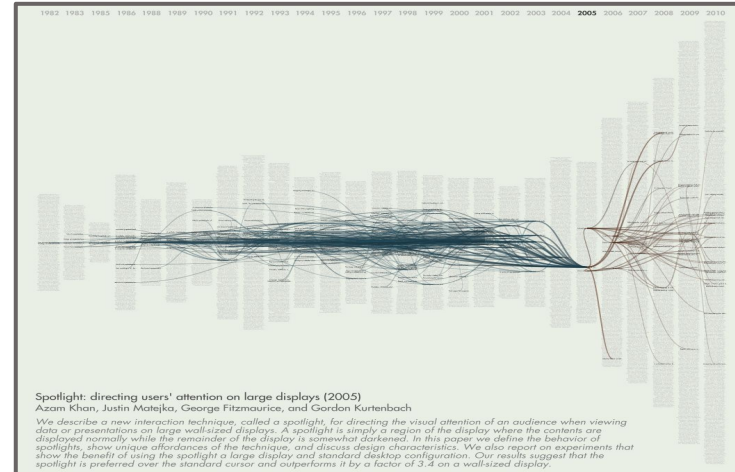
- **Deep learning approach**

- Complete sentence
- Attention model
- Pointer & Generator



Network Visualization

- Reference Network
 - Given one ETD, citation relationship between papers and ETDs is visualized
- Social Network
 - Collaboration strength between research groups
 - Author social network



Working on.....

- **User Study- based on the prospective Stakeholders**
 - Questionnaires to reach broader audience
 - Interview (Focus group)
 - Hands on use and feedback
- **Analysis of the collected data**
- **Presentation/Discussion of the result based on the analysis**



Reference

Sumit Bhatia, Cornelia Caragea, Hung-Hsuan Chen, Jian Wu, Pucktada Treeratpituk, Zhaohui Wu, Madian Khabza, Prasenjit Mitra, and C. Lee Giles. Specialized research datasets in the CiteSeerX digital library. *D-Lib Magazine*, 18(7/8), 2012.

Cornelia Caragea, Jian Wu, Alina Ciobanu, Kyle Williams, Juan Fernandez-Ramirez, Hung-Hsuan Chen, Zhaohui Wu, and Lee Giles. *CiteSeerX: A Scholarly Big Dataset*, pages 311–322. Springer International Publishing, Cham, 2014.

C. Lee Giles. The future of citeseer: Citeseerx. In *Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Germany, September 18-22, 2006, Proceedings, page 2, 2006.

Complete list of references: <https://goo.gl/QVgPQN>



Thank you!!!

