Creating Effective Training Datasets for Machine Learning

Part 1 – Zoning Documents for Data Extraction

Innodata®

TABLE OF CONTENTS

01 INTRODUCTION

As companies look for ways to reap the benefits of artificial intelligence and machine learning, they need as much data as feasibly possible to train and improve their models. The challenge lies in obtaining the right data, in the right formats and systems, and in the right quantity. After all, AI and machine learning will only be as smart as the quality of data that is collected. That said, there are several steps companies can employ on their journey to creating effective training datasets.

One such step is a process referred to as zoning. Zoning is specifically designed to turn unstructured information typically stored in documents like contracts and purchase orders into readily accessible smaller blocks of data that can be used in machine learning environments to drive smarter business outcomes.

These documents are typically ripe with valuable data points. For example, in the financial services industry, there are massive amounts of data generated every time a contract is created or amended. This information is always in flux, rapidly evolving during the typical lifespan of the agreement. Financial institutions must keep up with the changes to understand how this information may affect them. Therefore, they need to be able to seamlessly extract, analyze and manage critical data points within the document.

Documents Ripe For Zoning

- Contracts
- Press Releases
- Resumes
- Journals
- Research Reports
- Purchase Orders
- Shipping Information
- Invoices
- Investor Summaries
- Privacy Policies

But readily accessing and extracting this information has been historically difficult. These documents are typically in Portable Document Format – PDF. While PDF has become the digital equivalent of paper and the industry-standard for storing various types of information including everything from text to images, it is hard for computers to read and understand. The reality is it's next to impossible for a human to manually do all the work themselves. Therefore, the ability to transform these often complex documents into normalized, computer-addressable data is a critical step in developing training data.

Enter Zoning

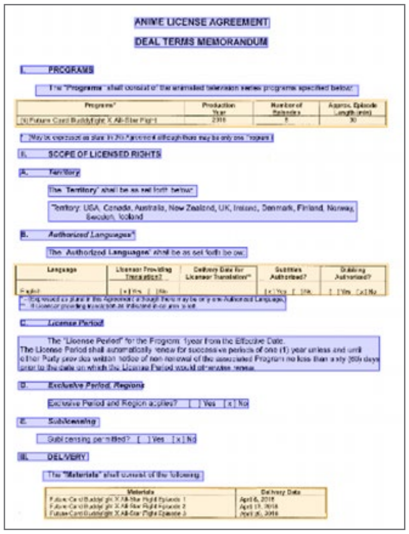
Breaking down PDF data into structured XML is a multi-step process. Zoning is the first step in the PDF conversion process and intends to automate the process of recognizing and classifying sequences and blocks of information within a document and then mapping it to a predefined "zone" category. This helps identify and categorize content in the PDF into different content types like abstract, title, image, references, authors etc. After categorization, these content blocks could be further processed through machine learning models to perform sequence labelling on them. This also presents the ability to perform complex tasks like extracting text from an image.

Without zoning, it would be a much more complex and expensive process to perform sequence labelling on specific blocks of text. Therefore, the ability to employ zoning on all content within PDF's and categorizing them into different zone types is an essential endeavour.

02 OUR APPROACH TOWARDS ZONING

The data pipeline for the extraction and structuring engine was conceptualized to perform the following functions:

- Input PDF
  ↓
- Zone PDF
  ↓
- Data Extraction
  ↓
- Annotate XML
  ↓
- Output XML



ANIME LICENSE AGREEMENT

DEAL TERMS MEMORANDUM

I. PROGRAMS

The "Programs" shall consist of the animated television series programs specified below:

Programs*	Production Year	Number of Episodes	Approx. Episode Length (min)
(1) Future Card Buddyfight X All-Star Fight	2018	6	30

* May be expressed as plural in this Agreement although there may be only one "Program."

II. SCOPE OF LICENSED RIGHTS

A. Territory

The "Territory" shall be as set forth below:

Territory: USA, Canada, Australia, New Zealand, UK, Ireland, Denmark, Finland, Norway, Sweden, Iceland

B. Authorized Languages*

The "Authorized Languages" shall be as set forth below:

Language	Licensor Providing Translation?	Delivery Date for Licensor Translation**	Subtitles Authorized?	Dubbing Authorized?
English	[x] Yes [ ] No		[x] Yes [ ] No	[ ] Yes [x] No

* Expressed as plural in this Agreement although there may be only one Authorized Language.
** If Licensor providing translation as indicated in column to left.

C. License Period

The "License Period" for the Program: 1year from the Effective Date.
The License Period shall automatically renew for successive periods of one (1) year unless and until either Party provides written notice of non-renewal of the associated Program no less than sixty (60) days prior to the date on which the License Period would otherwise renew.

D. Exclusive Period, Regions

Exclusive Period and Region applies? [ ] Yes [x] No

E. Sublicensing

Sublicensing permitted? [ ] Yes [x] No

III. DELIVERY

The "Materials" shall consist of the following:

Materials	Delivery Date
Future Card Buddyfight X All-Star Fight Episode 1	April 6, 2018
Future Card Buddyfight X All-Star Fight Episode 2	April 13, 2018
Future Card Buddyfight X All-Star Fight Episode 3	April 20, 2018

Innodata has been building a cognitive extraction and structuring engine to address the need of classifying and extracting complex documents. The platform accepts a PDF document as input, extracts information from the document, performs transformations and generates an annotated/tagged XML as output. We also employ a workbench where a user would create bound boxes around these content blocks within the PDF's. Coordinates from these bound boxes could be extracted. The content within these coordinates could be processed through an OCR (Optical Character Recognition) engine to get the text content.

Throughout this whitepaper, we'll go into more detail about the zoning process and share our learnings from our own journey with zoning.

Input PDF

The Input PDF microservice accepts a PDF in two formats, viz. PDF Normal and scanned PDF's. Normal PDF is converted into images (one per page).

Zone PDF

For each page image, the zoning microservice detects text and draws a box around it – which is called a zone. These zones are labelled as text, image, table or maths depending on the type of content in these zones. The zoning microservice uses Object Detection with Deep Learning to identify zones and their labels.

The adjacent screenshot is an example of a zoned document. These zones are then used for text extraction in subsequent microservices.

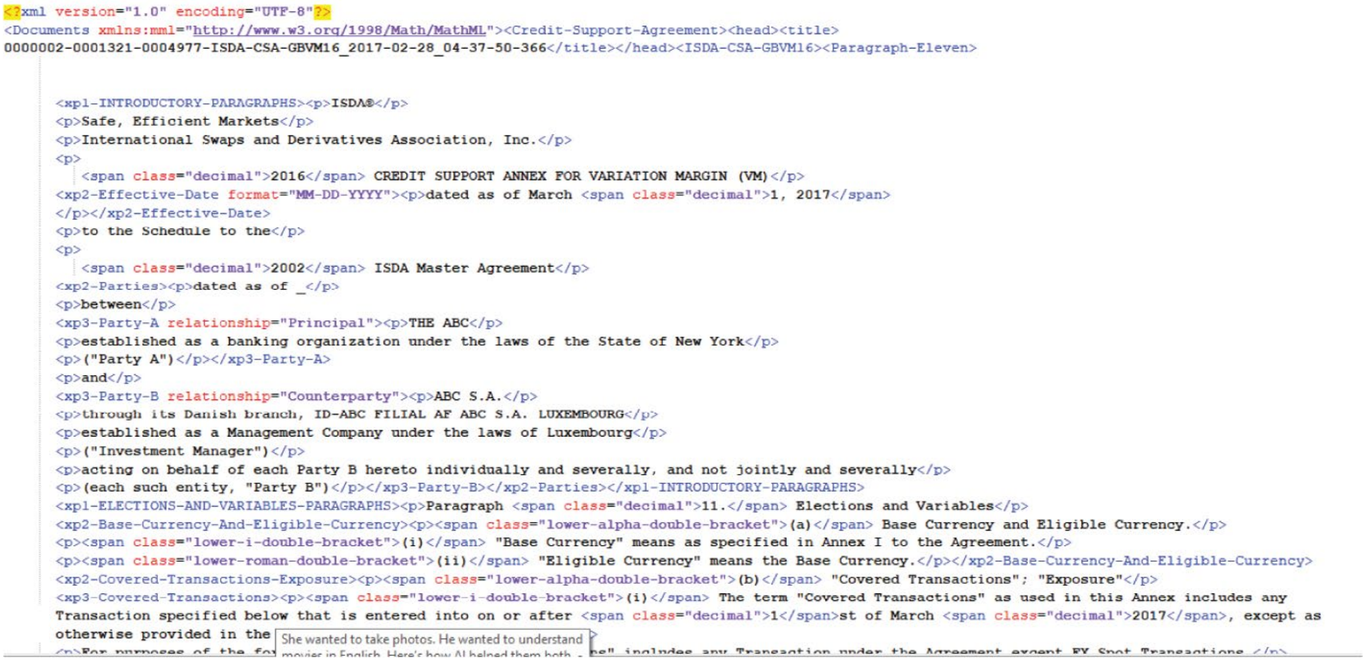
Data Extraction

Data extraction microservice takes a zoned document and extracts text from all the different zones. We use tesseract and other OCR tools for text extraction.

Annotate XML

The annotate XML microservice is powered by a deep learning TensorFlow model which performs sequence labelling on the extracted text. We'll go into more about annotation in part 2.

Here is a sample output from the annotate XML service:



```
<?xml version="1.0" encoding="UTF-8"?>
<Documents xmlns:mml="http://www.w3.org/1998/Math/MathML"><Credit-Support-Agreement><head><title>
0000002-0001321-0004977-ISDA-CSA-GBVM16_2017-02-28_04-37-50-366</title></head><ISDA-CSA-GBVM16><Paragraph-Eleven>


    <xp1-INTRODUCTORY-PARAGRAPHS><p>ISDA®</p>
    <p>Safe, Efficient Markets</p>
    <p>International Swaps and Derivatives Association, Inc.</p>
    <p>
      <span class="decimal">2016</span> CREDIT SUPPORT ANNEX FOR VARIATION MARGIN (VM)</p>
    <xp2-Effective-Date format="MM-DD-YYYY"><p>dated as of March <span class="decimal">1, 2017</span>
    </p></xp2-Effective-Date>
    <p>to the Schedule to the</p>
    <p>
      <span class="decimal">2002</span> ISDA Master Agreement</p>
    <xp2-Parties><p>dated as of _</p>
    <p>between</p>
    <xp3-Party-A relationship="Principal"><p>THE ABC</p>
    <p>established as a banking organization under the laws of the State of New York</p>
    <p>("Party A")</p></xp3-Party-A>
    <p>and</p>
    <xp3-Party-B relationship="Counterparty"><p>ABC S.A.</p>
    <p>through its Danish branch, ID-ABC FILIAL AF ABC S.A. LUXEMBOURG</p>
    <p>established as a Management Company under the laws of Luxembourg</p>
    <p>("Investment Manager")</p>
    <p>acting on behalf of each Party B hereto individually and severally, and not jointly and severally</p>
    <p>(each such entity, "Party B")</p></xp3-Party-B></xp2-Parties></xp1-INTRODUCTORY-PARAGRAPHS>
    <xp1-ELECTIONS-AND-VARIABLES-PARAGRAPHS><p>Paragraph <span class="decimal">11.</span> Elections and Variables</p>
    <xp2-Base-Currency-And-Eligible-Currency><p><span class="lower-alpha-double-bracket">(a)</span> Base Currency and Eligible Currency.</p>
    <p><span class="lower-i-double-bracket">(i)</span> "Base Currency" means as specified in Annex I to the Agreement.</p>
    <p><span class="lower-roman-double-bracket">(ii)</span> "Eligible Currency" means the Base Currency.</p></xp2-Base-Currency-And-Eligible-Currency>
    <xp2-Covered-Transactions-Exposure><p><span class="lower-alpha-double-bracket">(b)</span> "Covered Transactions"; "Exposure"</p>
    <xp3-Covered-Transactions><p><span class="lower-i-double-bracket">(i)</span> The term "Covered Transactions" as used in this Annex includes any
    Transaction specified below that is entered into on or after <span class="decimal">1</span>st of March <span class="decimal">2017</span>, except as
    otherwise provided in the She wanted to take photos. He wanted to understand >
    <p>For purposes of the fo movies in English. Here's how AI helped them both. - es" includes any Transaction under the Agreement except FX Spot Transactions.</p>
```

03 CHALLENGES – OUR ZONE DATASET

For the Zone PDF microservice, a key requirement was to identify relevant datasets. The dataset needed to be representative of the common production workloads and scenarios. This dataset also needed to have coordinate information of zones within the PDF document along with labels of the zones.

Iteration One - Data Collection

For building a microservice like Zone PDF, the key drivers were:

- Data availability and ease of collection (preferably within Innodata)
- Data sufficiency (enough number of samples)
- Representative of source files.

Our contract analysis and risk management platform, DocGenix, appeared to be an ideal candidate. The project data was easy to source as the data was available internally, and historical data for many years was also available with the needed document types (contract data).

The challenge was the complexity of DocGenix tools . The DocGenix applications were built over the years and several external plugins were incrementally added. Moreover, apart from its core tools, the team also used several external systems to do bits and pieces of the job functions.

Understanding Data

Our team spent several sessions spread over weeks to understand the DocGenix workflow.

The team discovered that DSRS and IETZ, internal Innodata applications, are used by DocGenix for zoning of PDF's. These applications work in tandem to identify zones within the DocGenix data (PDF). These zones are then processed using OCR (FR11) and subsequent manual review. The output of this process consists of two files- an image(tif) and an XML containing zone information for each block, line, word, and paragraph. These outputs are further processed by DocGenix teams to label and tag these zones.

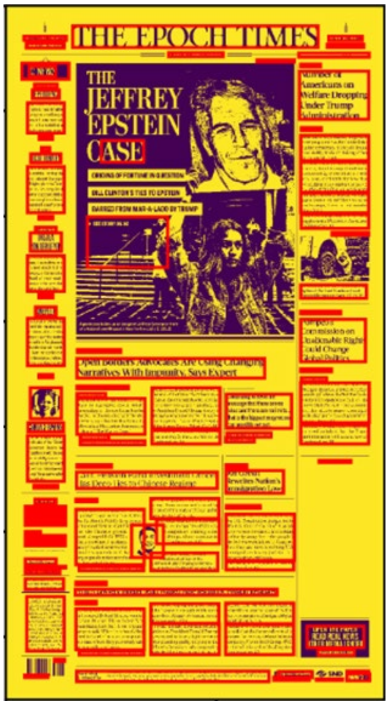
We spent about a week running from one team to another to understand this dataset and the subsequent steps. After numerous sessions with the teams, we realized that once this output is processed by the DocGenix teams, they strip out the zone (coordinate) information as it is not useful for the final deliverables. The transformation keeps only the content information and does away with the zone information.

Just when we thought we had our first dataset, this new piece of information sent us back to the drawing board.

Iteration Two: Search for Relevant Datasets

There are various other projects within Innodata which zone PDFs. We decided to reach out to other business units for zoned datasets even though we knew that we may not get zoned contracts data.

We spent the next 2-3 weeks talking to different teams within Innodata to understand their data.



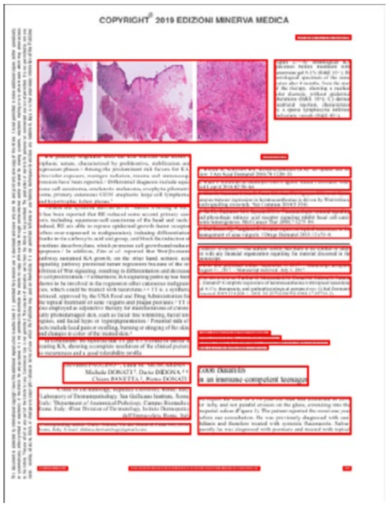
THE EPOCH TIMES

THE JEFFREY EPSTEIN CASE

Number of Americans on Welfare Dropping Under Trump Administration

Open Borders Advocates Are Using Changing Narratives With Impunity, Says Expert

Figure 1



COPYRIGHT© 2019 EDIZIONI MINERVA MEDICA

Figure 2

Invariably, every dataset that existed within Innodata didn't fit our needs. We could have used these datasets if we invested in fixing these datasets for our use. We realized that fixing an existing dataset would mean human capital investment, as we may possibly need a new set of tools and programs to be built.

We evaluated many projects and moved on to the next, hoping to find a more usable dataset. The intention was that if we do not find a completely usable dataset, we might just fix one of the existing datasets and make it usable for our project.

Take for example a project from one of our business units which had overlapping zones (Figure 1). We required clearly demarcated zones, making this project an unlikely candidate for our purpose.

Here is another example of a project that we evaluated (Figure 2). Though this did not have overlapping zones, the documents in this dataset had many unzoned parts. The production teams had zoned only the parts that they needed and left the others unzoned.

Yet in another project, the documents had complete zones, but images were not zoned. In one project, all images were zoned but had very little to no text.

There was one project which looked very promising; review of the first few samples suggested it had near perfect zones. However, upon further discussion we found that the zones were created by an in-house tool, and most zones were not reviewed and corrected by humans.

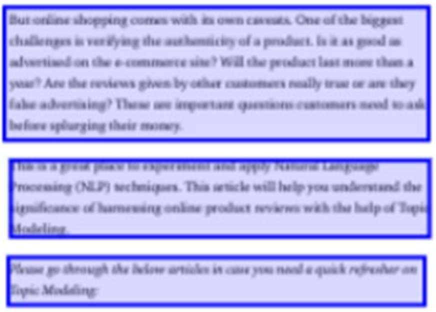
During these next 2-3 weeks we saw a plethora of datasets, every one missing some or more critical pieces.

04 THE EUREKA MOMENT - FINAL ITERATION

After spending weeks and repeated failures to get hold of a usable dataset for training, we decided to create a golden dataset – our own dataset which would conform to our requirements of both representation and relevance.

We had a zoning workbench which was basic but functional. We borrowed a few SMEs from the production BU and started training them on the kind of zoned datasets that we needed. We acquired contracts in PDF format from various projects within Innodata.

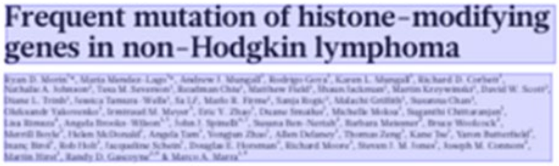
Knowing very well that we could manually zone limited data, we needed to be adamant about quality so as to ensure a quality dataset. We wanted to be as accurate as possible. We narrowed down our zone types to just 4 – text, image, table and math. We believed that these 4 zone types would be sufficient for our needs. We created a zoning instruction manual which gave an overview when and when not to choose a certain zone type.



But online shopping comes with its own caveats. One of the biggest challenges is verifying the authenticity of a product. Is it as good as advertised on the e-commerce site? Will the product last more than a year? Are the reviews given by other customers really true or are they false advertising? These are important questions customers need to ask before splurging their money.

This is a great place to experiment and apply Natural Language Processing (NLP) techniques. This article will help you understand the significance of harnessing online product reviews with the help of Topic Modeling.

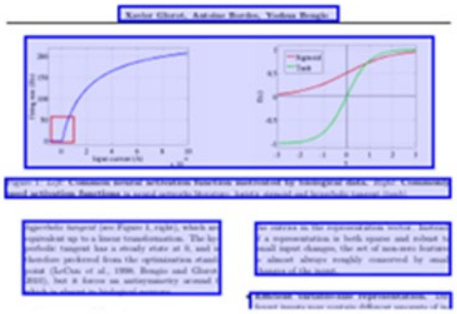
Please go through the below articles in case you need a quick refresher on Topic Modeling:



Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma

Ryan D. Morin, Maria Mendez-Lago, Andrew J. Mungall, Rodrigo Goya, Karen L. Mungall, Richard D. Corbett, Nathalie A. Johnson, Tesa M. Severson, Readman Chiu, Matthew Field, Shaun Jackman, Martin Krzywinski, David W. Scott, Diane L. Trinh, Jessica Tamura-Wells, Sa Li, Marlo R. Firme, Sanja Rogic, Malachi Griffith, Susanna Chan, Oleksandr Yakovenko, Irmtraud M. Meyer, Eric Y. Zhao, Duane Smailus, Michelle Moksa, Suganthi Chittaranjan, Lisa Rimsza, Angela Brooks-Wilson, John J. Spinelli, Susana Ben-Neriah, Barbara Meissner, Bruce Woolcock, Merrill Boyle, Helen McDonald, Angela Tam, Yongjun Zhao, Allen Delaney, Thomas Zeng, Kane Tse, Yaron Butterfield, Inanç Birol, Rob Holt, Jacqueline Schein, Douglas E. Horsman, Richard Moore, Steven J. M. Jones, Joseph M. Connors, Martin Hirst, Randy D. Gascoyne & Marco A. Marra

Zone Type - Text



Xavier Glorot, Antoine Bordes, Yoshua Bengio

Figure 1: *Left:* **Common neural activation function motivated by biological data.** *Right:* **Commonly used activation functions** in neural networks literature: logistic sigmoid and hyperbolic tangent (tanh).

hyperbolic tangent (see Figure 1, right), which are equivalent up to a linear transformation. The hyperbolic tangent has a steady state at 0, and is therefore preferred from the optimization standpoint (LeCun et al., 1998; Bengio and Glorot, 2010), but it forces an antisymmetry around 0 which is absent in biological neurons.

the entries in the representation vector. Instead, if a representation is both sparse and robust to small input changes, the set of non-zero features is almost always roughly conserved by small changes of the input.

Efficient variable-size representation. [illegible]

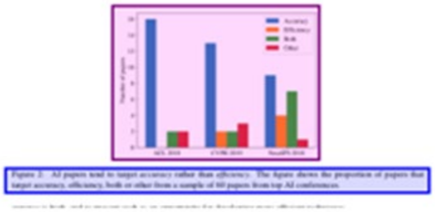
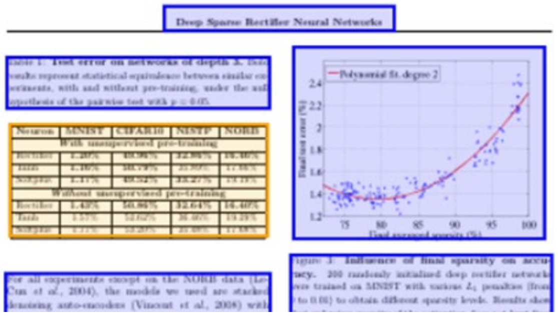


Figure 2: AI papers tend to target accuracy rather than efficiency. The figure shows the proportion of papers that target accuracy, efficiency, both or other from a sample of 60 papers from top AI conferences.

Zone Type - Image

Initially our zoning instruction document was small; we thought it was all sorted. However, with the passage of time we saw many different and varied types of data. We kept adding to this documentation and today this document is about 3 times the size of what we had started with. Only when you start working with the data do you realize that there are large number of variations in seemingly simple looking zone types – text, image, table and maths.



Deep Sparse Rectifier Neural Networks

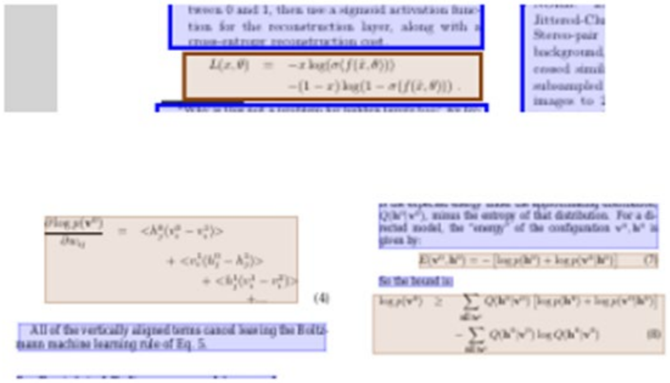
Table 1: **Test error on networks of depth 3.** Bold results represent statistical equivalence between similar experiments, with and without pre-training, under the null hypothesis of the pairwise test with $p = 0.05$.

Neuron	MNIST	CIFAR10	NISTP	NORB
With unsupervised pre-training				
Rectifier	1.20%	49.96%	32.86%	16.46%
Tanh	1.16%	50.79%	35.89%	17.66%
Softplus	1.17%	49.52%	33.27%	19.19%
Without unsupervised pre-training				
Rectifier	1.43%	50.86%	32.64%	16.40%
Tanh	1.57%	52.62%	36.46%	19.29%
Softplus	1.77%	53.20%	35.48%	17.68%

For all experiments except on the NORB data (LeCun et al., 2004), the models we used are stacked denoising auto-encoders (Vincent et al., 2008) with

Figure 3: **Influence of final sparsity on accuracy.** 200 randomly initialized deep rectifier networks were trained on MNIST with various L_1 penalties (from 0 to 0.01) to obtain different sparsity levels. Results show

Zone Type -Table



tween 0 and 1, then use a sigmoid activation function for the reconstruction layer, along with a cross-entropy reconstruction cost.

$$L(x,\theta) = -x\log(\sigma(f(\tilde{x},\theta))) - (1-x)\log(1-\sigma(f(\tilde{x},\theta))) .$$

Jittered-Cluttered NORB Stereo-pair background, crossed simil[illegible] subsampled images to [illegible]

$$\frac{\partial \log p(\mathbf{v}^0)}{\partial w_{ij}} = <h_j^0(v_i^0 - v_i^1)> + <v_i^1(h_j^0 - h_j^1)> + <h_j^1(v_i^1 - v_i^2)> + \ldots \quad (4)$$

All of the vertically aligned terms cancel leaving the Boltzmann machine learning rule of Eq. 5.

Q(h^0|v^0), minus the entropy of that distribution. For a directed model, the "energy" of the configuration v^0, h^0 is given by:

$$E(\mathbf{v}^0,\mathbf{h}^0) = -[\log p(\mathbf{h}^0) + \log p(\mathbf{v}^0|\mathbf{h}^0)] \quad (7)$$

So the bound is:

$$\log p(\mathbf{v}^0) \geq \sum_{\text{all } \mathbf{h}^0} Q(\mathbf{h}^0|\mathbf{v}^0)[\log p(\mathbf{h}^0) + \log p(\mathbf{v}^0|\mathbf{h}^0)] - \sum_{\text{all } \mathbf{h}^0} Q(\mathbf{h}^0|\mathbf{v}^0)\log Q(\mathbf{h}^0|\mathbf{v}^0) \quad (8)$$

Zone Type -Maths

05 CREATING THE PERFECT DATASET - PROCESS ENGINEERING

Since we were going to create manual datasets, we realized we could create only a small finite set- as creating a vast dataset would involve more time and more manual effort.

To start with, we assigned the same documents to all SMEs. We had instructed them to work in silos and not discuss and decide what kind of zones they would mark for content. In the end, we would review and match their zones to identify any differences in judgement. This gave us a different perspective to look at seemingly simple data.

In the next step, we gave the data zoned by one SME to another for review and correction. We encountered many new possibilities, even during these sessions. After this, we let the SMEs discuss among themselves and asked them to make the best collective judgement when in doubt.

Once we felt our zoning process was more stable, our SMEs started bulk zoning. We are just a few weeks into this process, and they still come to us with new scenarios which we discuss and classify into one or the other zone type.

06 NEXT STEPS FOR ZONING

We have a small first dataset, which is ready. This was used to train our zoning microservice. The initial results are quite promising.

Since we have the main zoning microservice in place, we can process all new documents through this microservice. Our SMEs are the humans-in-the-loop; they correct these documents manually. The effort of zoning has been considerably reduced since the microservice is doing a big chunk of the work and the SMEs are simply expected to correct and mark any incorrectly zoned elements.

Ultimately, we are now able to quickly work on complex documents and deliver the quality results our clients demand for training data.

We are simultaneously working on the next step – the annotation service. In part 2, we'll dive into this next critical component on the journey to creating effective datasets.

Key Things to Remember for Zoning:

- **Zoning is used primarily to identify an object within a larger number of objects.**
- **There are many use cases where zoning could be used - Face identification, people counting, self-driving cars, security, etc.**
- **While collecting training data, we should place special emphasis on the variety of data based on their presentation structure (e.g. the layout and zones in a magazine or newspaper would be different than contracts or scientific article.)**
- **The zones created in the training data need to be highly accurate because the machine learning models learn based on pixels information. Inaccurate zoning would lead to an inadequately trained model producing less than desirable results.**

Contact us:

www.innodata.com
email: info@innodata.com
Ph: 201-371-8000