Practical Systems-Level Development Challenges

Danna Gurari
University of Colorado Boulder
Spring 2024

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2024/
Review

• Last lecture
  • Multimodal applications
  • Image captioning dataset challenges
  • Image captioning algorithms
  • Visual question answering dataset challenges
  • Visual question answering algorithms
  • Foundation models

• Assignments (Canvas)
  • Final project outline due in 1.5 weeks

• Questions?
Today’s Topics

• Motivation

• Data curation

• Model maintenance

• TAs’ personal experiences
Today’s Topics

• Motivation
• Data curation
• Model maintenance
• TAs’ personal experiences
Most Code for a DL System Is Not for DL

Sculley et al. Neurips 2015
Frequently Asked Questions in Deep Learning

Source: 39,628 Stack Overflow questions from a dump in December 2018 with one of these tags: tensorflow, pytorch, or deeplearning4j

Approach: human labeling for initial set followed by automated labeling
Much Research Examines Efforts to Build Software with AI Capabilities; e.g.,

(Neurips 2015)

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley, gholt, dgg, edavydov, toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner, vchaudhary, mwyoung, jfcrespo, dennison}@google.com
Google, Inc.
Much Research Examines Efforts to Build Software with AI Capabilities; e.g.,

Software Engineering for Machine Learning: A Case Study

Saleema Amershi  
Microsoft Research  
Redmond, WA USA  
samershi@microsoft.com

Andrew Begel  
Microsoft Research  
Redmond, WA USA  
andrew.begel@microsoft.com

Christian Bird  
Microsoft Research  
Redmond, WA USA  
cbird@microsoft.com

Robert DeLine  
Microsoft Research  
Redmond, WA USA  
rdeline@microsoft.com

Harald Gall  
University of Zurich  
Zurich, Switzerland  
gall@ifi.uzh.ch

Ece Kamar  
Microsoft Research  
Redmond, WA USA  
eckamar@microsoft.com

Nachiappan Nagappan  
Microsoft Research  
Redmond, WA USA  
nachin@microsoft.com

Besmira Nushi  
Microsoft Research  
Redmond, WA USA  
besmira.nushi@microsoft.com

Thomas Zimmermann  
Microsoft Research  
Redmond, WA USA  
tzimmer@microsoft.com
Much Research Examines Efforts to Build Software with AI Capabilities; e.g.,

Common ML Workflow (from Microsoft Study of Software Teams Building Software Applications with AI Capabilities)
Key Theme: Like Code, **Data Is a Core Part of Software Engineering**
Today’s Topics

• Motivation
• Data curation
• Model maintenance
• TAs’ personal experiences
“Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI

Nithya Sambasivan
nithyasamba@google.com
Google Research
Mountain View, CA

Shivani Kapania
kapania@google.com
Google Research
Mountain View, CA

Hannah Highfill
hhighfil@google.com
Google Research
Mountain View, CA

Diana Akrong
dakrong@google.com
Google Research
Mountain View, CA

Praveen Paritosh
pkp@google.com
Google Research
Mountain View, CA

Lora Aroyo
loraa@google.com
Google Research
Mountain View, CA
Data Curation

Amershi et al. ICSE-SEIP 2019
Key Approaches

- Internet-curated
- User labeled
Key Approaches

• Internet-curated

• User labeled
e.g., Recall Common Crawl

Common Crawl is a 501(c)(3) non-profit founded in 2007.

We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

Overview

Over 250 billion pages spanning 17 years.
Free and open corpus since 2007.
Cited in over 10,000 research papers.
3–5 billion new pages added each month.

Popular NLP source (web archives without HTML markup and non-text content [https://commoncrawl.org/](https://commoncrawl.org/))

Many variants exist with different filtering
e.g., Colossal Clean Crawled Corpus (C4)

April 2019 Web Snap Shot, filtering including by:

- only retaining lines ended in a terminal punctuation mark (e.g. a period, exclamation mark, question mark)
- discarding pages with less than 3 sentences and lines with less than 5 words
- discarding pages with any word on the “List of Dirty, Naughty, Obscene or Otherwise Bad Words”.
- discarding pages with the phrase “lorem ipsum”
Curating Image-Text Pairs from Wikipedia; e.g.,

1. Image-Text Pair Collection

- Source: Wikipedia, given its high quality (editorially reviewed), large size (~124M pages), and diversity (279 languages)
- Extracted ~150 million image-text pairs

Berimbau

The berimbau (Portuguese pronunciation: [beɾimboj], borrowed from Kimbundu mbirimbu[1]) is a traditional Angolan musical bow that is commonly used in Brazil.[2]

It consists of a single-stringed bow attached to a gourd resonator and is played with a stick and a coin or stone to create different tones and rhythms.

The berimbau was used in many parts of Africa and Brazil during the 19th century to accompany chants and storytelling.[2] It is part of the candomblé tradition, later incorporated into the Afro-Brazilian art capoeira. Until the mid-20th century, it was used almost exclusively within the black community, but after the popularization of capoeira, it gain wider popularity.

Today, berimbau is used in various genres of popular music.

History  [edit]

Berimbau is an adaptation of African gourde musical bows, as no Indigenous Brazilian or European people use musical bows.[2][3] According to the musicologist Gerard Kubik, the berimbau and the *southwest Angolan variety called mbulumbumba are identical in construction and playing technique, as well as in tuning and in a number of basic patterns played.[4] The assimilation of this Angolan instrument is evident also in other Bantu terms used for musical bow in Brazilian Portuguese, including urucungo, and madimba linguungu.

In 1859, French journalist Charles Ribeyrolles described free practices of African slaves on

Angola musical bow (1922), known as berimbau in Brazil.
Curating Image-Text Pairs from Wikipedia; e.g.,

1. Image-Text Pair Collection
   - Source: Wikipedia, given its high quality (editorially reviewed), large size (~124M pages), and diversity (279 languages)
   - Extracted ~150 million image-text pairs

2. Filtering
   - Removed images with "generic" or meaningless text (e.g., maps), unsuitable licenses, questionable content (e.g., pornography, violence), and width or height < 100 pixels
   - Only kept example in top 100 languages

3. Human Quality Validation
   - Crowdsourced ratings for nearly 4,400 examples
   - Majority vote label used from 3 independent ratings
   - Examples were in English (~3,000), German (300), French (300), Spanish (300), Russian (300), Chinese (300), & Hindi (100)

Results from first two questions suggested both reference and attribution texts are high-quality.

No major difference found across different languages.
Curating Image-Text Pairs from **Wikipedia**; e.g.,

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Images</th>
<th>Text</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr30K [39]</td>
<td>32K</td>
<td>158K</td>
<td>&lt; 8</td>
</tr>
<tr>
<td>SBU Captions [24]</td>
<td>~1M</td>
<td>~1M</td>
<td>1</td>
</tr>
<tr>
<td>MS-COCO [21]</td>
<td>~330K</td>
<td>~1.5M</td>
<td>&lt; 4</td>
</tr>
<tr>
<td>CC [5]</td>
<td>~3.3M</td>
<td>~3.3M</td>
<td>1</td>
</tr>
<tr>
<td><strong>WIT</strong></td>
<td>11.5M</td>
<td>37.6M</td>
<td>108</td>
</tr>
</tbody>
</table>

WIT has 37.6 million (image, text) pairs describing 11.5 million unique images spanning 108 languages (each with 12K+ examples)

Key Approaches

• Internet-curated

• User labeled
Typical Workflow

1. Category Selection
   - Choose categories of interest

2. Data Collection
   - Downloaded relevant pages from Internet

(Adapted from slides by Antonio Torralba)
Typical Workflow

1. Category Selection
   Choose categories of interest

2. Data Collection
   Downloaded relevant pages from Internet

3. Human Labeling/Verification
   Humans manually inspect and possibly label data to establish final labels
Key Approaches

• Internet-curated

• User labeled
Key Considerations

• Suitable license? e.g.,

https://teaching.resources.osu.edu/teaching-topics/simple-guide-creative-commons
Key Considerations

• Suitable license? e.g., ImageNet redacted

The rise and fall (and rise) of datasets

Growing criticisms of datasets that were built from user-generated data scraped from the web have led to the retirement or redaction of many popular benchmarks. Their afterlife, as copies or subsets that continue to be used, is a cause for concern.
Key Considerations

• Suitable license? e.g., Common Crawl ended curation of copyrighted content linked to by millions of URLs

Key Considerations: What Else?

• Suitable license?

• Storage costs?

• Efficient visualization/search tools?

• Data scale (e.g., how to collect rare content; e.g., medical/satellite-based)

• Data distribution (e.g., what’s the skew of representation across contents)
Today’s Topics

• Motivation

• Data curation

• Model maintenance

• TAs’ personal experiences
Model Maintenance

Amershi et al. ICSE-SEIP 2019
Often need rapid release cycles of new models to support current trends, given how quickly data can become outdated; e.g., evolving
• shopping products
• politics (e.g., who is current president)
• building/product designs (e.g., for blind people’s visual interpretation services)
Testing Frameworks

Data suites often are created to test hard cases and/or a model’s “ethics”

Unique Software Engineering Requirements

• Unlike traditional software development:
  
  • Modular design is challenging for DL components (we often choose DL because it is impractical to identify human-based rules)

  • Data management and versioning is critical
Today’s Topics

• Motivation

• Data curation

• Model maintenance

• TAs’ personal experiences
How do you spend your time when developing DL methods? (and other questions from students)

Everley

Nick
Today’s Topics

- Motivation
- Data curation
- Model maintenance
- TAs’ personal experiences
The End