Machine Learning (ML): A New Tool for Libraries & Archives

Bohyun Kim
Associate University Librarian for Library Information Technology,
University of Michigan, USA
Twitter: @bohyunkim | Web: http://www.bohyunkim.net/blog/

Aeolian Network - Online Workshop 4: AI/ML: Increasing Access, Visibility, and Engagement, April 20, 2022
Example of 2D semantic segmentation: (Top) input image (Bottom) prediction.
Supervised learning

input data

□ □ ... □
(images)

labels
dog, cat ... dog

ML
model

prediction
this is a cat

new UNSEEN image

Unsupervised learning

input data

□ □ ... □
(images)

ML
model

groupings
cat 1, cat 2
dog 1, dog 2

bird 1, bird 2

Reinforcement learning

environment

action

state

reward

interpreter

Agent

A mostly complete chart of Neural Networks

Human Generated Data

Title
Seated Bodhisattva Avalokitesvara (Guanyin Pusa), One from a Pair of Bodhisattvas, Probably from an Amitabha Triad

Date
581 – 618

Classification
Sculpture

Credit Line
Harvard Art Museums/Arthur M. Sackler Museum, Bequest of Grenville L. Winthrop, 1943.53.26

Machine Generated Data

Tags

<table>
<thead>
<tr>
<th>Amazon</th>
<th>Clarifai</th>
<th>Imagga</th>
<th>Google</th>
<th>Microsoft</th>
</tr>
</thead>
<tbody>
<tr>
<td>created on 2019-04-07</td>
<td>created on 2018-03-16</td>
<td>created on 2018-03-16</td>
<td>created on 2018-03-16</td>
<td>created on 2018-03-16</td>
</tr>
<tr>
<td>Worship 96.3</td>
<td>sculpture 100</td>
<td>ruler 96.5</td>
<td>sculpture 94.8</td>
<td>building 88.6</td>
</tr>
<tr>
<td>Buddha 93.9</td>
<td>art 99.7</td>
<td>statue 67.8</td>
<td>statue 87.1</td>
<td>sculpture 82</td>
</tr>
<tr>
<td>Art 93.9</td>
<td>statue 99.3</td>
<td>temple 63.9</td>
<td>classical sculpture 85.7</td>
<td>statue 51.4</td>
</tr>
<tr>
<td>Human 89.7</td>
<td>religion 98.8</td>
<td>religion 63.8</td>
<td>gautama buddha 78.5</td>
<td></td>
</tr>
<tr>
<td>Person 85.1</td>
<td>ancient 97</td>
<td>sculpture 56</td>
<td>ancient history 73.8</td>
<td></td>
</tr>
<tr>
<td>Figurine 59.4</td>
<td>god 96.8</td>
<td>culture 49.7</td>
<td>bronze 68.9</td>
<td></td>
</tr>
<tr>
<td>Sculpture 59</td>
<td>Buddha 96.8</td>
<td>ancient 39.9</td>
<td>artifact 67.7</td>
<td></td>
</tr>
</tbody>
</table>
Can ML be a useful tool for libraries/archives?

Figure 10. Excerpt from Herman B Wells facial recognition contact sheet.

Q. What would a practical application of ML for libraries/archives look like?
Current Applications of Machine Learning at Libraries/Archives

I. ML for expediting image processing/analysis
II. ML for expanding descriptive metadata
III. ML for enhancing metadata
I. ML for Expediting Archival Image Processing/Analysis

“Processing” refers to the arrangement, description, and housing of materials for storage and use.

Unprocessed materials pose an impediment to researchers because they cannot be easily discovered, accessed, and used.

Machine learning can expedite the processing of archival materials.
Image Analysis for Archival Discovery (Aida)

The Aida research team explores applications of image analysis and machine learning in digital libraries of historic materials. We’re especially interested in what we might learn from the millions of digital images that librarians, archivists, and others are creating as they digitize the cultural record. We’re intrigued by the questions that machine learning approaches might help to surface in these collections and about our professional practices—and also by the questions our collections and professional practices might help to surface about machine learning.

Our current and recent efforts include “Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project” (Library of Congress), “Extending Image Analysis for Archival Discovery” (IMLS, LG-71-16-0152-16), and “Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914” (subaward on IMLS, LG-00-17-0104-17).

Code & Data

Code developed for our project is available via our GitHub organization page. See also the Exploring-ML-with-Project-Aida repository distributed by the Library of Congress.

Data generated for our project are made available through appropriate data repositories designed for long-term storage, preservation, and access. See our project page on the Open Science Framework: https://osf.io/xn7tv/
The AIDA Project Explored:

- Document Segmentation;
- Graphic Element Classification & Text Extraction;
- Document Type Classification;
- Digitization Type Differentiation;
- Document Image Quality Assessment; and
- Document Clustering.

(a) Identify/Classify Graphical Content & Pull Text with Image Zoning and Segmentation

• The AIDA team tried to identify and classify graphical content (e.g. figures, illustrations, and cartoons) and extract text in historical newspaper page images, using ML’s image zoning and segmentation (U-NeXt).

• 2 sets of newspaper images were used for ML model training/evaluation.
  • One set from the Library of Congress’s Beyond Words project, which crowdsourced the graphical content zone demarcation and the text transcription of historical newspaper images.
    • 1,532 images and corresponding data from graphical content zones; 1,226 images used for training and 306 for evaluation.
  • The other set from the Europeana Newspapers Project.
    • A set of 481 page images already zoned and segmented, with the segments classed as background, text, figure, separator or table.
What is Computer Vision (CV)?

• Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs — and take actions or make recommendations based on that information.
Computer Vision Concepts

- Image filtering, matching, segmentation, and recognition
- Segmentation divides whole images into pixel groupings which can then be labelled and classified.

Images from: https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html
Tasks Performed by Computer Vision

- Image classification
- Object detection
- Object tracking
- Content-based image retrieval
- Semantic segmentation
- Instance segmentation

Image from: https://labelbox.com/image-segmentation-overview
(b) Natural Image / Document Classification

• The AIDA project team applied natural image and document classification methods in machine learning to a collection of minimally processed manuscript materials to see whether handwritten, printed, and mixed (both handwritten and printed) documents can be distinguished from one another.

• The team generated their own model by using the VGG method with 16 categories (VGG-16) pre-trained on the RVL_CDIP dataset.

• 999 images were used in total.

<table>
<thead>
<tr>
<th></th>
<th>Handwritten</th>
<th>Typed/Typescript</th>
<th>Mixed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td>267</td>
<td>267</td>
<td>267</td>
<td>801</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>99</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>99</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>999</td>
</tr>
</tbody>
</table>
Figure 9. Prediction failure cases. In the left example, the model classified the document as handwritten rather than Mixed. Note that the printed regions are very small compared to the handwritten content in the image. In the right example, the model classified the document as printed rather than mixed. Here, the handwritten region is very small compared to the printed region in the image.

(c) Detecting Digitization Type

• The AIDA project team also applied machine learning techniques to both classifying manuscript collection images as digitized from original vs. from microform and analyzing their image quality.

• 36,103 images were used as the ground truth classification

<table>
<thead>
<tr>
<th>Total Images</th>
<th>Expected Microform Source</th>
<th>Classified Microform Source</th>
<th>Expected Original Source</th>
<th>Classified Original Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>36,103</td>
<td>2,256</td>
<td>2,834</td>
<td>33,847</td>
<td>33,269</td>
</tr>
</tbody>
</table>

Figure 13. A digital image of a photograph, digitized from the original item, and classified by the classifier as having been digitized from a microform source.

Source: Elizabeth Lorang et al., “Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project” (University of Nebraska - Lincoln, January 10, 2020), [https://digitalcommons.unl.edu/libraryscience/396](https://digitalcommons.unl.edu/libraryscience/396).
Benefits of ML for Archival Processing

• There is no doubt that being able to identify graphical content contained in archival materials and to peruse the text in them would make browsing and navigating those materials a lot more manageable to users. And if ML can process archival materials this way, it would surely improve their discoverability and usefulness.

• Automating the classification of materials as handwritten, printed, and mixed can surely save archivists’ and librarians’ time spent on performing such classification manually.

• Automatically assessing the quality of images and assigning their original source type can help archivists and librarians identify materials at higher-risk and take appropriate preservation measures.

Two Challenges

• The results from these explorations indicate that these machine learning methods are not ready to be immediately put to use in archival processing.
  • “Training a classifier to learn information from rare classes is very hard.”

• The project report highlights two challenges:
  • (a) general scarcity of training data sets;
  • (b) overall difficulty in obtaining or establishing ground truth (= the expected result against which the performance of a machine learning model is measured).

Need for More Data and Ground Truth

• Without a sufficiently large amount of training data, a machine learning model for libraries/archives cannot be created.

• Without ground-truth datasets, it is difficult to tell how well a machine learning model for libraries/archives works.
II. ML for Expanding Descriptive Metadata

Mass digitization projects created many digital objects of historical materials, and many more born-digital items are being added to archival collections.

Unfortunately, libraries and archives lack resources to manually create rich metadata for all digital objects in a timely manner.

Lacking such metadata, those digital objects are not easily discovered, identified, or navigated.

ML can advance the discovery, navigation, and clustering of these materials by generating more detailed metadata.
(a) Image Classifier ML Algorithm for the Frick Collection

• The Frick Art Reference Library in New York together with Stanford University, Cornell University, and the University of Toronto developed an image classifier algorithm that applies a local classification system based on visual elements to the library’s digitized Photoarchive.

• Tens of thousands of already-labeled images & hundreds of thousands of images that require labeling.

• Hierarchical classification headings:
  • “American School: Portraits: Men: Without hands (Without hats): Head to left”
Process

• The Cornell/Toronto/Stanford team modified the popular ResNet152 network—with 152 layers and six million parameters—and trained it on a collection of 45,857 classified reproductions of American paintings provided by the Frick Art Reference Library’s Photoarchive.

• The majority were portraits and represented 98 unique classification headings.

• After training, the network was fed images from the unlabeled portion of the library’s Photoarchive, and the network predicted a classification heading for each one.

Vetting & Prospect

• These images were then annotated with the predictions and shown to the library’s photoarchivists through an application to vet the predictions.

• The library staff vetted 8,661 images in the year August 2019 - August 2020 and agreed with the network on the classification headings applied to 5,829 (67 %) of the images.

• Yet even the incorrect predictions were, in general, “almost correct” with only one tag incorrect or missing.

• Holds the potential to improve the staff’s efficiency in in metadata creation and image retrieval.
(b) CAMPI (Computer-Aided Metadata Generation for Photo archives Initiative)
Expanding Item-Level Description Metadata

• The CAMPI project team built a prototype application that uses computer vision (CV), in order to aid archivists and metadata editors for Carnegie Mellon University Archives’ General Photograph Collection (GPC) of +20,000 images with metadata at some hierarchy (e.g. job; directory).

• This collection of photographic prints, 35mm negatives, and born digital images was specifically chosen for the project, because:
  • those images lack item-level description and
  • are stored in and accessed through a shared network drive by archivists and metadata editors, with no browsing interface.

Visual Similarity Search

• The CAMPI prototype application offered archivists and metadata editors visual similarity search using CV techniques in ML.

• For this visual similarity search capability, a pre-trained “off-the-shelf” image search model was used.

• The team used the second-to-last pooled layer of the InceptionV3 trained neural network as the feature source. For each image, this produced an array of 512 numbers, allowing them to treat the entire collection of images as a set of points in 512-dimensional space.

• Then, the team used an approximate nearest neighbor search algorithm to create an index of that feature space, allowing it to be queried.
The prototype allows staff to search similar images to reduce repeated metadata generation and remove noise from the eventual public discovery interface.

Figure 1: A visualization of the photographs from this collection in the 512-dimensional feature space produced by the InceptionV3 neural network. Points have been projected into two dimensional space using UMAP. Photos are colored based on automated k-means clustering\(^7\), and we have manually labeled a handful solely to give a general sense of how the feature space corresponds to different visual content & attributes of the photos—these categories and exact clusters are not part of our actual prototype.
Speeding Up Tagging

• The application provided a browsing and CV-assisted tagging interface faceted by the existing job and directory metadata associated with the photographs.

• Using this application, archivists and metadata editors were able to start with a tag, find a seed photo, and retrieve similar photos via visual similarity search that specifies the number of neighbors requested.

• Then, they were able to quickly tag those retrieved photos with tags from the seed photo when appropriate.
Figure 10. Tagging workflow

1. Editors select a tag to work on

2. Editors use the browse interface to select an initial seed photograph
ML for Metadata Workflow

• In their white paper, the CAMPI team reported that of the more than 43,000 tagging decisions made, a little over 1 in 5 were able to be automatically distributed across a set, saving metadata editors more than 9,000 decisions.

• They also noted that the prototype application identified around 28% of the collection as sets with duplicates.

• This prototype had the effect of rapidly streamlining and speeding up item-level metadata creation.

• These results demonstrate how machine learning can be integrated into the existing metadata creation and editing workflow at libraries and archives.

• What is interesting about the CAMPI project is that even though the use of computer vision was minimal in their prototype application, this project produced a successful outcome because it combined machine learning with a well-designed user interface that took advantage of the existing archival structure, thereby effectively augmenting archivists’ and metadata experts’ existing work.

• At libraries and archives, a browsing and filtering interface is primarily created to provide patrons with public access to archival materials.

• However, such an interface significantly increases the overall productivity of metadata-related work, if it fits in well with the way the staff perform their work.
Need for More Training Sets

• “There is a field-wide need for specialized computer vision training sets based on historical, nonborn-digital photograph collections.”

III. ML for Enhancing Metadata

The Audiovisual Metadata Platform Pilot Development (AMPPD) Project is run by Indiana University Libraries, University of Texas at Austin School of Information, New York Public Library, and AVP, a data management solutions and consulting company from late 2018 to June 2021.

The goal was to generate and manage metadata for audiovisual materials at scale using ML, by creating automated metadata generation mechanism (MGM)s and integrating them with the human metadata generation process.

The final project report was released in Dec 2021.
AMP: Audiovisual Metadata Platform

AMP, or Audiovisual Metadata Platform, is an open source software system, currently under development, that enables the efficient generation of metadata to support discovery and use of digitized and born-digital audio and moving image collections. In July 2021, the Indiana University (IU) Libraries, in collaboration with New York Public Library and digital consultant AVP, were awarded a new grant from the Andrew W. Mellon Foundation to support development work to make AMP more easily deployable and usable and to pilot test AMP using collections at IU and NYPL.

The original development of AMP from 2018-2021 was supported by a previous Mellon Foundation grant. That grant was preceded by a workshop and resulting white paper funded by the Mellon Foundation and hosted by IU as part of a planning project. The partners leading this planning project were the IU Libraries, University of Texas at Austin (UT) School of Information, and AVP.

In the years leading up to the planning project workshop, the project partners had embarked upon various initiatives investigating audiovisual description. In 2015, IU and AVP investigated models and developed a strategy for high-throughput description of audiovisual materials that are being digitized as part of IU’s Media Digitization and Preservation Initiative (MDPI).

**Project Background Information**

- AMPPD Final Project Report, December 2021
- AMP (Phase 3) Mellon Grant Proposal Narrative (2021-2022)
- AMPPD (Phase 2) Mellon Grant Proposal Narrative (2018-2021)
- Audiovisual Metadata Platform (AMP) Planning Project: Progress Report and Next Steps. 2018. Dunn, Jon W.; Hardesty, Juliet L.; Clement, Tanya; Lacinak,
Generating Rich Metadata for Audiovisual Materials at Scale

• AMPPD Project’s Priority Tasks
  • First: Detecting silence, speech, and music; speech-to-text / speaker diarization; named entity recognition, such as people, geographic locations, and topics;
  • Next: Video and structured data OCR (metadata, transcript, subject terms); detecting scene and shot; detecting music genre in audiovisual materials.

Project Outcome

• The AMPDD project built a fully functioning platform, integrating 24 MGMs that could be used to analyze, describe, and document the audiovisual materials in IU and NYPL’s collections, using both proprietary and open-source ML tools.

• Human MGMs for transcript correction and named entity revision were also fully incorporated in AMP.

ML for Better Discovery, Identification, and Navigation of AV Materials

• The team had greater success with more common machine-learning problems, such as speech-to-text transcription and named entity recognition.

• By contrast, music genre classification and extracting structured information on performers and works performed, by running zoning OCR on printed music programs, proved to be more difficult.
  • The MGM team also successfully trained custom models for two categories: applause detection and facial recognition (of known specific individuals).

• This pilot is a big step towards AV materials’ better discovery, identification, and navigation.
Future Focuses for Libraries/Archives

• Need for more data & ground truth for ML algorithm training and evaluation

• Exploration of pre-trained off-the-shelf ML models
  • e.g. AWS Rekognition (aws.amazon.com/rekognition) and Google Cloud Vision AI (cloud.google.com/vision).

• Need for continued digitization
Thank you!

Bohyun Kim
Associate University Librarian for Library Information Technology, University of Michigan, USA
Twitter: @bohyunkim | Web: http://www.bohyunkim.net/blog/