Chapter 11

Taking a Leap Forward: Machine Learning for New Limits

Patrice-Andre Prud'homme
Oklahoma State University

Introduction

Today, machines can analyze vast amounts of data and increasingly produce accurate results through the repetition of mathematical or computational procedures. With the increasing computing capabilities available to us today, artificial intelligence (AI) and machine applications have made a leap forward. These rapid technological changes are inevitably influencing our interpretation of what AI can do and how it can affect people's lives. Machine learning models that are developed on the basis of statistical patterns from observed data provide new opportunities to augment our knowledge of text, photographs, and other types of data in support of research and education. However, "the viability of machine learning and artificial intelligence is predicated on the representativeness and quality of the data that they are trained on," as Thomas Padilla, Interim Head, Knowledge Production at the University of Nevada Las Vegas, asserts (2019, 14). With that in mind, these technologies and methodologies could help augment the capacity of archives and libraries to leverage their creation-value and minimize their institutional memory loss while enhancing the interdisciplinary approach to research and scholarship.

In this essay, I begin by placing artificial intelligence and machine learning in context, then proceed by discussing why AI matters for archives and libraries, and describing the techniques used in a pilot automation project from the perspective of digital curation at Oklahoma State University Archives. Lastly, I end by challenging other areas in the library and adjacent fields to join in the dialogue, to develop a machine learning solution more broadly, and to explore opportunities that we can reap by reaching out to others who share a similar interest in connecting people to build knowledge.

Artificial Intelligence and Machine Learning. Why do they Matter?

Artificial intelligence has seen a resurging interest in the recent past—in the news, in the literature, in academic libraries and archives, and in other fields, such as medical imaging, inspection of steel corrosion, and more. John McCarthy, American computer scientist, defined artificial intelligence as "the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable" (2007, 2). This definition has since been extended to reflect a deeper understanding of AI today and what systems run by computers are now able to do. Dr. Carmel Kent notes that "AI feels like a moving target" as we still need to learn how it affects our lives (2019). Within the last decades, the amazing jump in computing capabilities has been quite transformative in that machines are increasingly able to ingest and analyze large amounts of data and more complex data to automatically produce models that can deliver faster and more accurate results.[1] Their "power lies in the fact that machines can recognize patterns efficiently and routinely, at a scale and speed that humans cannot approach," writes Catherine Nicole Coleman, digital research architect for Stanford University (2017).

A Paradigm Shift for Archives and Libraries

Within the context of university archives, this paradigm shift has been transforming the way we interpret archival data. Artificial intelligence, and specifically machine learning as a subfield of AI, has direct applications through pattern recognition techniques that predict the labeling values for unlabeled data. As the software analytics company SAS argues, it is "the iterative aspect of machine learning [that] is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results" (n.d.).

Case in point, how can we use machine learning to train machines and apply facial and text recognition techniques to interpret the sheer number of photographs and texts in either analog or born-digital formats held in archives and libraries? Combining automatic processes to assist in supporting inventory management with a focus on descriptive metadata, a machine learning solution could help alleviate time-consuming and relatively expensive metadata tagging tasks, and thus scale the process more effectively using relatively small amounts of data. However, the traditional approach of machine learning would still require a significant time commitment by archivists and curators to identify essential features to make patterns usable for data training. By contrast, deep learning algorithms are able "to learn high-level features from data in an incremental manner. This eliminates the need of domain expertise and hard core feature extraction" (Mahapatra 2018).

Deep learning has regained popularity since the mid-2000s due to "fast development of high-performance parallel computing systems, such as GPU clusters" (Zhao 2019, 3213). Deep learning neural networks are more effective in feature detection as they are able to solve complex problems such as image classification with greater accuracy when trained with large datasets. The challenge is whether archives and libraries can afford to take advantage of greater computing capabilities to develop sophisticated techniques and make complex patterns from thousands of

[1] See SAS n.d. and Brennan 2019.

digital works. The sheer size of library and archive datasets, such as university photograph collections, presents challenges to properly using these new, sophisticated techniques. As Jason Griffey writes, "AI is only as good as its training data and the weighting that is given to the system as it learns to make decisions. If that data is biased, contains bad examples of decision-making, or is simply collected in such a way that it isn't representative of the entirety of the problem set[...], that system is going to produce broken, biased, and bad outputs" (2019, 8). How can cultural heritage institutions ensure that their machine learning algorithms avoid such bad outputs?

Implications to Machine Learning

Machine learning has the potential to enrich the value of digital collections by building upon experts' knowledge. It can also help identify resources that archivists and curators may never have the time for, and at the same time correct assumptions about heritage materials. It can generate the necessary added value to support the mission of archives and libraries in providing a public good. Annie Schweikert states that "artificial intelligence and machine learning tools are considered by many to be the next step in streamlining workflows and easing workloads" (2019, 6).

For images, how can archives build a data-labeling pipeline into their digital curation workflow that enables machine learning of collections? With the objective being to augment knowledge and create value, how can archives and libraries "bring the skills and knowledge of library staff, scholars, and students together to design an intelligent information system" (Coleman 2017)? Despite the opportunities to augment knowledge from facial recognition, models generated by machine learning algorithms should be scrutinized so long it is unclear how choices are made in feature selection. Machine learning "has the potential to reveal things ...that we did not know and did not want to know" as Charlie Harper asserts (2018). It can also have direct ethical implications, leading to biased interpretations for nefarious motives.

Machine Learning and Deep Learning on the Grounds of Generating Value

In the fall 2018, Oklahoma State University Archives began to look more closely at a machine learning solution to facilitate metadata creation in support of curation, preservation, and discovery. Conceptually, we envisioned boosting the curation of digital assets, setting up policies to prioritize digital preservation and access for education and research, and enhancing the long-term value of those data. In this section, I describe the parameters of automation and machine learning used to support inventory work and experiment with face recognition models to add contextualization to digital objects. From a digital curation perspective, the objective is to explore ways to add value to digital objects for which little information is known, if any, in order to increase the visibility of archival collections.

What started this Pilot Project?

Before proceeding, we needed to gain a deeper understanding of the large quantity of files held in the archives—both types of data and metadata. The challenge was that with so many files, so many formats, files become duplicated and renamed, doctored, and scattered throughout directories to accommodate different types of projects over time, making it hard to sift due to sparse

metadata tags that may have differed from one system to another. In short, how could we justify the value of these digital assets for curatorial purposes? How much could we rely on the established institutional memory within the archives? Lastly, could machine learning or deep learning applications help us build a greater capacity to augment knowledge? In order to optimize resources and systematically make sense of data, we needed to determine that machine learning could generate value, which in turn could help us more tightly integrate our digital initiatives with machine learning applications. Such applications would only be as effective as the data are good for training and the value we could derive from them.

Methodology and Plan of Action

First, we recruited two student interns to create a series of processes that would automatically populate a comprehensive inventory of all digital collections, including finding duplicate files by hashing. We generated the inventory by developing a process that could be universally adapted to all library digital collections, setting up a universal list of works and their associated metadata, with a focus on descriptive metadata, which in turn could support digital curation and discovery of archival materials—digitized analog materials and born-digital materials. We developed a universal policy for digital archival collections, which would allow us to incorporate all forms of metadata into a single format to remedy inconsistencies in existing metadata. This first phase was critical in the sense that it would condition the cleansing and organizing of data. We could then proceed with the design of a face recognition database, with the intent to trace individuals featured in the inventory works of the archives to the extent that our data were accurate. We utilized the Oklahoma State University Yearbook collections and other digital collections as authoritative references for other works, for the purpose of contextualization to augment our data capacity.

Second, we implemented our plan; worked closely with the Library Systems' team within a Windows-based environment; decided on Graphics Processing Unit (GPU) performance and cost, taking into consideration that training neural networks necessitates computing power; determined storage needs; and fulfilled other logistical requirements to begin the step-by-step process of establishing a pattern recognition database. We designed the database on known objects before introducing and comparing new data to contextualize each entry. With this framework, we would be able to add general metadata tags to a uniform storage system using deep learning technology.

Third, we applied Tesseract OCR on a series of archival image-text combinations from the archives to extract printed text from those images and photographs. "Tesseract 4 adds a new neural net (LSTM) [Long Short-Term Memory] based OCR engine which is focused on line recognition," while also recognizing character patterns ("Tesseract" n.d.). We were able to obtain successful output for the most part, with the exception of a few characters that were hard to detect due to pixelation and font types.

Fourth, we looked into object identifiers, keeping in mind that "When there are scarce or insufficient labeled data, pre-training is usually conducted" (Zhao 2019, 3215). Working through the inventory process, we knew that we would also need to label more data to grow our capacity. We chose to use ResNet 50, a smaller version backbone of Keras-Retinanet, frequently used as a starting point for transfer learning. ResNet 152 was another implementation layer used as shown in Figure 11.1 demonstrating the output of a training session or epoch for testing purposes.

Keras is a deep learning network API (Application Programming Interface) that supports multiple back-end neural network computation engines (Heller 2019) and RetinaNet is a sin-

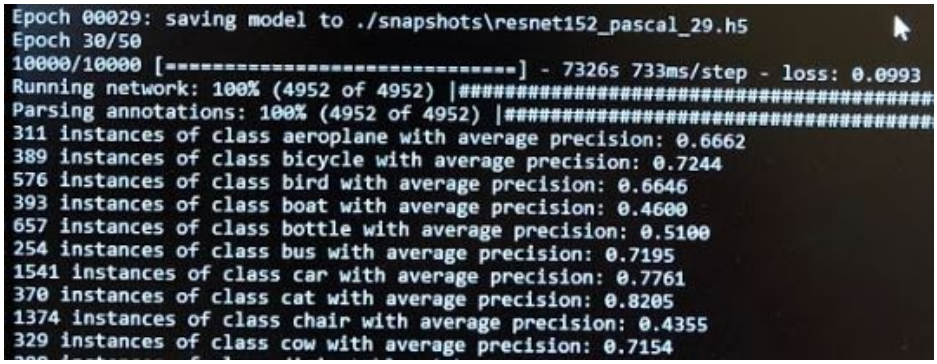

Figure 11.1: ResNet 152 application using PASCAL VOC 2012

Figure 11.2: Face recognition API test

gle, unified network consisting of a backbone network and two task-specific subnetworks used for object detection (Karaka 2019). We proceeded by first dumping a lot of pre-tagged information from pre-existing datasets into this neural network. We experimented with three open source datasets: PASCAL VOC 2012, a set including 20 object categories; Open Images Database (OID), a very large dataset annotated with image-level labels and object bounding boxes; and Microsoft COCO, a large-scale object detection, segmentation, and captioning dataset. With a few faces from the OID dataset, we could compare and see if a face was previously recognized. Expanding our process to data known from the archives collection, we determined facial areas, and more specifically, assigned bounding box regressions to feed into the facial recognition API, based on Keras code written in Python. The face recognition API is available via GitHub. [2] It uses a method called Histogram of Oriented Gradient (HOG) encoding that makes the actual face recognition process much easier to implement for individuals because the encodings are fairly unique for every person, as opposed to encoding images and trying to blindly figure out which parts are faces based on our label boxes. Figure 11.2 illustrates our test, confirming from two very different photographs the presence of Jessie Thatcher Bost, the first female graduate from Oklahoma A&M College in 1897.

Ren et al. stated that it is important to construct a deep and convolutional per-region object

[2] See `https://github.com/ageitgey/face_recognition`.

classifier to obtain good accuracy using ResNets (2015). Going forward, we could use the tool "as is" despite the low tolerance for accuracy, or instead try to establish large datasets of faces by training on our own collections in hopes of improving accuracy. We proceeded with utilizing the Oklahoma State University Yearbook collections, comparing image sets with other photographs that may include these faces. We look forward to automating more of these processes.

A Conclusive First Experiment

We can say that our first experiment developing a machine learning solution on a known set of archival data resulted in positive output, while recognizing that it is still a work in progress. For example, the model we ran for the pilot is not natively supported on Windows, which hindered team collaboration. In light of these challenges, we think that our experiment was a step in the right direction of adding value to collections by bringing in a new layer of discovery for hidden or unidentified content.

Above all, this type of work relies greatly on transparency. As Schweikert notes, "Transparency is not a perk, but a key to the responsible adoption of machine learning solutions" (2019, 72). More broadly, issues in transparency and ethics in machine learning are important concerns in the collecting and handling of data. In order to boost adoption and get more buy-in with this new type of discovery layer, our team shared information intentionally about the process to help add credibility to the work and foster a more collaborative environment within the library. Also, the team developed a Graphic User Interface (GUI) to search the inventory within the archives and ultimately grow the solution beyond the department.

Challenges and Opportunities of Machine Learning

Challenges

In a National Library of Medicine blog post, Patti Brennan points out "that AI applications are only as good as the data upon which they are trained and built"(2019), and having these data ready for analysis is a must in order to yield accurate results. Scaling of input and output variables also plays an important role in the performance improvement when using neural network models. Jerome Pesenti, Head of AI at Facebook, states that "When you scale deep learning, it tends to behave better and to be able to solve a broader task in a better way" (2019). Clifford Lynch affirms, "machine learning applications could substantially help archives make their collections more discoverable to the public, to the extent that memory organizations can develop the skills and workflows to apply them" (2019). This raises the question whether archives can also afford to create the large amount of data from print heritage materials or refine their born-digital collections in order to build the capacity to sustain the use of deep-learning applications. Granted, the increasing volume of born-digital materials could help leverage this data capacity somehow; it does not exclude the fact that all data will need to be ready prior to using deep learning. Since machine learning is only good so long as value is added, archives and libraries will need to think in terms of optimization as well, deciding when value-generated output is justified compared to the cost of computing infrastructure and skilled labor needs. Besides value, operations, such as storing and ensuring access to these data, are just as important considerations to making machine learning a feasible endeavor.

Opportunities

Investment in resources is also needed for interpreting results, in that "results of an AI-powered analysis should only factor into the final decision; they should not be the final arbiter of that decision" (Brennan 2019). While this could be a challenge in itself, it can also be an opportunity when machine learning helps minimize institutional memory loss in archives and libraries (e.g., when long-time archivists and librarians leave the institution). Machine learning could supplement practices that are already in place—it may not necessarily replace people—and at the same time generate metadata for the access and discovery of collections that people may never have the time to get to otherwise. But we will still need to determine accuracy in results. As deep learning applications will only be as effective as the data, archives and libraries should expand their capacity by working with academic departments and partnering with university supercomputing centers or other highly performant computing environments across consortium aggregating networks. Such networks provide a computing environment with greater data capacity and more GPUs. Along similar lines, there are opportunities to build upon Carpentries workshops and the communities of practice that surround this type of interest.

These growing opportunities could help boost the use of machine learning and deep learning applications to minimize our knowledge gaps about local history and the surrounding community, bringing together different types of data scattered across organizations. This increased capacity for knowledge could grow through collaborative partnerships, connecting people, scholars, computer scientists, archivists and librarians, to share their expertise through different types of projects. Such projects could emphasize the multi- and interdisciplinary academic approach to research, including digital humanities and other forms or models of digital scholarship.

Conclusion

Along with greater computing capabilities, artificial intelligence could be an opportunity for libraries and archives to boost the discovery of their digital collections by pushing text and image recognition machine learning techniques to new limits. Machine learning applications could help increase our knowledge of texts, photographs, and more, and determine their relevance within the context of research and education. It could minimize institutional memory loss, especially as long-time professionals are leaving the profession. However, these applications will only be as effective as the data are good for training and for the added value they generate.

At Oklahoma State University, we took a leap forward developing a machine learning solution to facilitate metadata creation in support of curation, preservation, and discovery. Our experiment with text extraction and face recognition models generated conclusive results within one academic year with two student interns. The team was satisfied with the final output and so was the library as we reported on our work. Again, it is still a work in progress and we look forward to taking another leap forward.

In sum, it will be organizations' responsibility to build their data capacity to sustain deep learning applications and justify their commitment of resources. Nonetheless, as Oklahoma State University's face recognition initiative suggests, these applications can augment archives' and libraries' support for multi- and interdisciplinary research and scholarship.

References

Brennan, Patti. 2019. "AI is Coming. Are Data Ready?" NLM Musings from the Mezzanine (blog). March 26, 2019. https://nlmdirector.nlm.nih.gov/2019/03/26/ai-is-coming-are-the-data-ready/.

Carmel, Kent. 2019. "Evidence Summary: Artificial Intelligence in Education." European EdTech Network. https://eetn.eu/knowledge/detail/Evidence-Summary-\/-Artificial-Intelligence-in-education.

Coleman, Catherine Nicole. 2017. "Artificial Intelligence and the Library of the Future, Revisited." Stanford Libraries (blog). November 3, 2017. https://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited.

"Face Recognition." n.d. Accessed November 30, 2019. https://github.com/ageitgey/face_recognition.

Griffey, Jason, ed.. 2019. "Artificial Intelligence and Machine Learning in Libraries." Special issue, *Library Technology Reports* 55, no. 1 (January). https://journals.ala.org/index.php/ltr/issue/viewIssue/709/471.

Harper, Charlie. 2018. "Machine Learning and the Library or: How I Learned to Stop Worrying and Love My Robot Overlords." *Code4Lib*, no. 41 (August). https://journal.code4lib.org/articles/13671.

Heller, Martin. 2019. "What is Keras? The Deep Neural Network API Explained." InfoWorld (website). January 28, 2019. https://www.infoworld.com/article/3336192/what-is-keras-the-deep-neural-network-api-explained.html.

Karaka, Anil. 2019. "Object Detection with RetinaNet." Weights & Biases (website). July 18, 2019. https://www.wandb.com/articles/object-detection-with-retinanet.

Lynch, Clifford. 2019. "Machine Learning, Archives and Special Collections: A High Level View." International Council on Archives Blog. October 1, 2019. https://blog-ica.org/2019/10/02/machine-learning-archives-and-special-collections-a-high-level-view/.

Mahapatra, Sambit. "Why Deep Learning over Traditional Machine Learning?" Towards Data Science (website). March 21, 2018. https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063.

McCarthy, John. "What is Artificial Intelligence?" Professor John McCarthy (website). Revised November 12, 2007. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf.

Padilla, Thomas. 2019. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries.* Dublin, OH: OCLC Research. https://doi.org/10.25333/xk7z-9g97.

Pesenti, Jerome. 2019. "Facebook's Head of AI Says the Field Will Soon 'Hit the Wall.'" Interview by Will Knight. Wired (website). December 4, 2019. https://www.wired.com/story/facebooks-ai-says-field-hit-wall/.

Ren, Shaoqing, Kaiming He, Ross Girshick, Xiangyu Zhang, and Jian Sun. 2015. "Object Detection Networks on Convolutional Feature Maps." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, no. 7 (April).

SAS. n.d. "Machine Learning: What It Is and Why It Matters." Accessed December 17, 2019.

https://www.sas.com/en_us/insights/analytics/machine-learning.html.

Schweikert, Annie. 2019. "Audiovisual Algorithms, New Techniques for Digital Processing." Master's Thesis, New York University. https://www.nyu.edu/tisch/preservation/program/student_work/2019spring/19s_thesis_Schweikert.pdf.

"Tesseract OCR." n.d. Accessed December 11, 2019. https://github.com/tesseract-ocr/tesseract.

Zhao, Zhong-Qiu, Peng Zheng, Shou-tao Xu, and Xindong Wu. 2017 "Object Detection with Deep Learning: A Review." *IEEE Transactions on Neural Networks and Learning Systems* 30, no. 11 (2019): 3212-3232.