Scalable Content-Based Analysis of Images in Web Archives with TensorFlow and the Archives Unleashed Toolkit

Hsiu-Wei Yang,[1] Linqing Liu,[1] Ian Milligan,[2] Nick Ruest,[3] and Jimmy Lin[1]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Department of History, University of Waterloo
[3] York University Libraries

ABSTRACT

We demonstrate the integration of the Archives Unleashed Toolkit, a scalable platform for exploring web archives, with Google's TensorFlow deep learning toolkit to provide scholars with content-based image analysis capabilities. By applying pretrained deep neural networks for object detection, we are able to extract images of common objects from a 4TB web archive of GeoCities, which we then compile into browsable collages. This case study illustrates the types of interesting analyses enabled by combining big data and deep learning capabilities.

1 INTRODUCTION

Efforts to systematically capture and preserve the web date back to the 1990s and the value of web archiving as an important form of cultural heritage preservation has been broadly recognized. However, tools to provide scholarly access have lagged, which is an acknowledged challenge in the community. To address this, researchers have begun to examine ways to help humanities scholars and social scientists work with web archives at scale. One particularly promising thread is the development of analytics toolkits leveraging "big data" platforms such as ArchiveSpark [2], Warcbase [3] and its successor, the Archives Unleashed Toolkit.

While these efforts have made substantial inroads into tackling the access challenge, one major shortcoming of most previous efforts is their focus on textual content—specifically, HTML. This makes sense as a starting point, but the web is a multimedia medium and web archives are no different. While analyses based on text, such as extracting and aggregating named-entities from HTML pages [3] are no doubt useful, the percentage of text on web pages peaked in 2005 and has been falling ever since; non-textual media is increasingly important [1]. While there have been previous explorations of web archive images via metadata and color clustering [6], we are not aware of systematic support in existing toolkits.

As a first step in addressing this gap, we demonstrate the integration of the Archives Unleashed Toolkit with neural network models for object recognition in Google's TensorFlow deep learning toolkit. This combination allows scholars to directly peer into the *content* of images in web archives at scale, which can augment image analysis based on filenames, the keyword context of the images, etc. Using pretrained object detection models from TensorFlow's "Model Zoo", we are able to automatically identify instances of dozens of different types of common objects, ranging from people to buses. With post-processing, we are able to create browsable collages of these images, with metadata and pointers back to their sources.

The contribution of this work is, to our knowledge, the first integration of deep learning models with a toolkit for exploring web archives to support content-based image analysis at scale. While we only present a simple case study here, this integration gives scholars entrée into the rich world of neural networks and deep learning technology for a multitude of image processing tasks. The flexibility of the Archives Unleashed Toolkit enables the exploration of web archives in ever more interesting ways, combining text, graphs, and now, image analysis.

2 IMPLEMENTATION

The Archives Unleashed Toolkit[1] (AUT) is best described as a Scala domain-specific language on top of the Apache Spark open-source data analysis platform, where users manipulate large web archives by defining data-parallel transformations over collections of records (called Resilient Distributed Datasets, or RDDs). Since TensorFlow[2] is built around Python, our first integration task was to bridge the two different programming languages. Fortunately, via PySpark (Python bindings for Spark), we are able to manipulate RDDs directly in Python.

After preliminary evaluation in terms of flexibility, speed, and accuracy, we decided to use the Single Shot MultiBox Detector [4] model available in TensorFlow. An important consideration in our application is inference latency, since humanities scholars are unlikely to have access to large compute clusters, and the large size of many web archives can lead to daunting end-to-end processing times. In our implementation, we broadcast the pretrained model to all Spark executors so that they can load model parameters and run inference in parallel to detect image content (i.e., with the `map` transformation). The output of the model is a list of objects that are detected in the image and associated probabilities.

Note that in our integration, inference with neural network models is treated like any user-defined function (UDF), just like, for example, a named-entity extractor. This allows image processing capabilities to be integrated into any other analysis using AUT: the output of the model can be subsequently filtered, aggregated, and manipulated just like any other RDD.

3 CASE STUDY

As a case study, we used the GeoCities collection provided by the Internet Archive to explore image extraction and analysis at scale. GeoCities was a web hosting platform founded in 1994 and closed in 2009; it had approximately seven million users and our collection consists of approximately 186M HTML pages [5]. The entire web archive totals 4TB. GeoCities is significant from a scholarly perspective because it encapsulates early web history—many users created their first website on the platform. We can imagine several types of research questions that can be explored through image analysis.

[1] https://github.com/archivesunleashed/aut/
[2] https://www.tensorflow.org/

978-1-7281-1547-4/19/$31.00 ©2019 IEEE
DOI 10.1109/JCDL.2019.00107

Figure 1: Screenshot of a collage of cars from GeoCities, overlaid with a zoomed-in portion (red box).

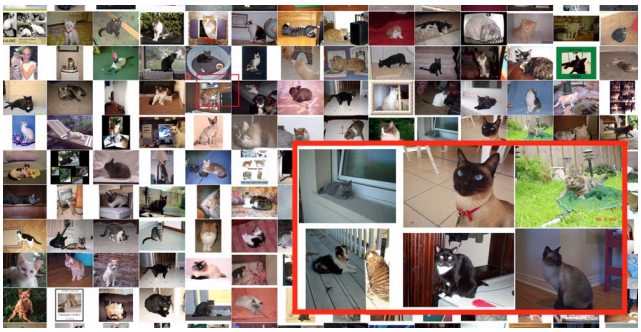

Figure 2: Screenshot of a collage of cats from GeoCities, overlaid with a zoomed-in portion (red box).

For example, GeoCities was (rather uniquely) arranged in thematic clusters called "neighborhoods", such as sites about philosophy in "Athens", cars in "MotorCity", and pets in "Heartland". Using object detection, we can find clusters of images that can suggest the existence of coherent communities. Did websites that were "near" each other in virtual space use similar images?

As another example, we can use a person detector to quickly understand the demographic characteristics of a collection. This is reminiscent of a famous digital humanities project, the *Invisible Australians* project, which used face detection to underpin the human scale of Australian immigration exclusion policy.[3] We note, however, that there are ethical considerations in performing such analyses at scale, but such discussions are beyond the scope of this paper (nevertheless, our community needs to have this discussion sooner rather than later). From a methodological perspective, through these inquiries we can exploit image analysis to counterbalance the dominance of text in digital humanities research.

Figure 1 shows a screenshot from a browsable collage comprising approximately 3000 cars from the GeoCities web archive:[4] our interface provides smooth pan and zoom capabilities to facilitate exploration. In this particular interface, we only consider images larger than 640×640 so we obtain a reasonably high-quality product. The collage was created by taking the output of our object detector, gathering the identified images, and feeding them to jutxa,[5] a utility that generates tile-based collages with metadata, rendered with OpenSeadragon. We can imagine a historian using this output as the starting point for a scholarly inquiry—for example, examining the association of images with particular GeoCities neighborhoods (such as, in this case, "MotorCity"). Similarly, with an eye to showcasing the model's versatility, we present cats in Figure 2.

We end with a rough performance analysis to characterize the computing resources that would be necessary to support image analysis. While GPUs greatly accelerate inference, we evaluated performance on the CPU since they are ubiquitous. There are approximately 2.3M images (larger than 640 × 640) in our GeoCities archive, and inference on a single image takes approximately 550ms. While there is a plethora of options in terms of model quality/cost tradeoffs, this provides a rough back-of-the-envelope estimate. We are able to analyze the entire collection in about a week on a *single* high-end server. This processing time can be greatly reduced with GPU-based inference (if such resources are available), but even on the CPU, we believe that image processing capabilities are within the reach of most humanities scholars.

[3] http://invisibleaustralians.org/
[4] https://ruebot.net/geocities-jcdl2019/
[5] https://github.com/tokee/juxta/

4 CONCLUSIONS

Lin et al. [3] proposed a process model for scholarly interactions with web archives that begins with a question and proceeds iteratively through four main steps: filter, analyze, aggregate, and visualize. What we've demonstrated here is an impoverished realization of this process: we perform no filtering since we process the entire archive; "analyze" involves applying inference with our neural network model; "aggregate" and "visualize" involve generating the output collage. Nevertheless, integration of TensorFlow and the Archives Unleashed Toolkit enables far more interesting analyses, potentially combining image analysis with existing capabilities—for example, enabling questions that simultaneously interrogate hyperlink structures, textual content, as well as image content. With these capabilities, the richness and depth of potential inquiries is only limited by the scholar's imagination and curiosity.

Acknowledgments. This work was primarily supported by the Natural Sciences and Engineering Research Council of Canada. Additional funding for this project has come from the Andrew W. Mellon Foundation. Our sincerest thanks to the Internet Archive for providing us with the GeoCities web archive.

REFERENCES

[1] A. Cocciolo. 2015. The Rise and Fall of Text on the Web: A Quantitative Study of Web Archives. *Information Research* 20, 3 (2015).
[2] H. Holzmann, V. Goel, and A. Anand. 2016. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *JCDL*. 83–92.
[3] J. Lin, I. Milligan, J. Wiebe, and A. Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *J. Comput. Cult. Herit.* 10, 4, Article 22 (July 2017), 30 pages.
[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. Berg. 2016. SSD: Single Shot MultiBox Detector. In *ECCV*. 21–37.
[5] I. Milligan. 2019. GeoCities. In *SAGE Handbook of Web History*, Niels Brügger and Ian Milligan (Eds.). SAGE Publications, London.
[6] I. Milligan. 2019. Learning to See the Past at Scale: Exploring Web Archives through Hundreds of Thousands of Images. In *Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision for History*, K. Kee and T. Compeau (Eds.). University of Michigan Press, Ann Arbor.