Content-Based Exploration of Archival Images Using Neural Networks

Tobi Adewoye,[1] Xiao Han,[1] Nick Ruest,[2] Ian Milligan,[3] Samantha Fritz,[3] Jimmy Lin[1]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] York University Libraries [3] Department of History, University of Waterloo

ABSTRACT

We present DAIRE (Deep Archival Image Retrieval Engine), an image exploration tool based on latent representations derived from neural networks, which allows scholars to "query" using an image of interest to rapidly find related images within a web archive. This work represents one part of our broader effort to move away from text-centric analyses of web archives and scholarly tools that are direct reflections of methods for accessing the live web. This short piece describes the implementation of our system and a case study on a subset of the GeoCities web archive.

ACM Reference Format:
Tobi Adewoye, Xiao Han, Nick Ruest, Ian Milligan, Samantha Fritz, and Jimmy Lin. 2020. Content-Based Exploration of Archival Images Using Neural Networks. In *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398577

1 INTRODUCTION

Methods for accessing web archives generally mimic methods for accessing the live web itself. For example, the Internet Archive's Wayback Machine and other comparable temporal browsing services allow users to specify a URL and navigate links to contemporaneous pages. Search capabilities in web archives are still nascent, but existing systems such as the Wayback Machine and Warclight [12] are largely modeled after existing web search engines: the user types a keyword query into a search box and gets back a list of results. Web archives, obviously, are not simply collections of textual documents, and image search has long been identified as an important capability desired by users [5]. However, there has been relatively little exploration along these lines [6], and more generally, on image analysis in web archives. Furthermore, we wonder if the most popular implementation of image search today—type in a few keywords, get back a gallery of images—is the most appropriate way to deploy image analysis capabilities on web archives.

In this short piece, we present an alternative, in a prototype called DAIRE (Deep Archival Image Retrieval Engine). Our system can be best characterized as a content-based image similarity exploration tool based on latent representations from neural networks. The "query" is an image from a web archive, and the results are other images from the same archive; there is no explicit keyword query, and the user can navigate endlessly by clicking on a result image and using *it* as the new query. We situate our tool in the broader context of web archival research and explain its design, presenting a case study on part of the GeoCities web archive.

2 THE VISUAL TURN

Most current efforts in building tools for analyzing web archives—including our own [8, 11]—have focused on webgraphs and textual content. The web is, of course, comprised of multiple media. Drawing on a sample of archived websites between 1999 and 2014, for example, one research study found that in 1999 the percentage of text on an average webpage was only around 22%; peaking in 2005 at 32% and declining to around 25% by 2014 [4]. Elsewhere in the digital humanities, scholars are increasingly attuned to the "visual turn": the need to not only "distantly read" material, but to "distantly view" them as well [1]; examples include applications of neural networks to analyze image collections [13]. Our own previous work has attempted to provide scholars with image access into web archives by taking advantage of object detectors based on neural networks to create collages that portray a multitude of pre-defined objects [14]. However, the general problem of how to best provide "distant viewing" tools remains far from solved.

We questioned whether a traditional keyword-based image search engine was the best tool for web archive exploration. Keywords are often insufficient to capture what scholars are interested in—for example, a search for images using the keyword "bears" could bring up images of wildlife, Winnie the Pooh, or even Chicago Bears football players. Modern image search engines group results into clusters to address this diversity, but here we explore an alternative approach. In the course of examining an image, a scholar may wish to find similar images—and also examine the pages in which those images appear. Are these similar images displayed in the same context? Are they part of a broader virtual community? Here, similarity is not necessarily defined in terms of objects contained within—and hence the limited usefulness of keyword search. When considering an image of Winnie the Pooh, images of other denizens of the Hundred Acre Wood might be of interest, as well as cartoons of a similar graphical style; but perhaps not, for example, wildlife photographs. These nuances are often difficult to convey via keywords, but easy to recognize visually if we provide well-designed exploration tools.

3 CASE STUDY: GEOCITIES

As a case study, we turned to the GeoCities collection provided by the Internet Archive, made available to us via a research agreement. 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 186 million HTML pages, totaling around

JCDL '20, August 1–5, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7585-6/20/08.
https://doi.org/10.1145/3383583.3398577

4TB of WARCs. GeoCities was arranged in thematic clusters called "neighborhoods", such as sites about philosophy in "Athens", cars in "MotorCity", and pets in "Heartland" [10].

We focused our case study on the "EnchantedForest": a GeoCities neighborhood that was designed to be around "pages by kids, for kids". In other words, the authors of these pages were likely kids or families, and a heavily-empowered team of moderators enforced strict content guidelines. This is of particular interest for two main reasons: First, one of the co-authors (Milligan) is a historian of childhood and youth; being able to work with user-generated content directly is fascinating, although it does raise significant ethical questions [7]. Second, the moderated nature of this neighborhood means that most of the images should be "safe for work" (although, as our team would discover, moderation is imperfect). As we hope to share our tool publicly, we wished to keep it free of pornography and other offensive material as much as possible.

Data preparation was carried out using the Archives Unleashed Toolkit [11], which grew out of our earlier efforts [8]. We took advantage of the recently-introduced Toolkit capabilities around Spark DataFrames, which present relational views of certain aspects of the collection and allow easy relational manipulations of the images, webpages, and image links (selection, projection, joins, etc.). We first enumerated all unique images in the "EnchantedForest" neighborhood based on MD5 hashes, recognizing that the same exact image could have different filenames on different webpages—as authors download and re-upload to share images, renaming them in the process. Indeed, this phenomenon itself is of interest to scholars since it captures a trace of how a community might have grown. In total, we were able to identify 369k images, 131k of which were unique. Via a join with the DataFrame capturing image links from the collection, we were able to identify the page that contains each image, which allows a scholar to navigate back to the original source and examine the context.

To facilitate content-based image exploration, we exploited the expressive power of convolutional neural networks to generate a compact representation of each image. After some preliminary evaluation, we decided to use the Xception V1 model [3] pretrained on ImageNet, which is available in Keras. Following the method of Babenko et al. [2], we performed inference on each of the 131k unique images and extracted the final pooling layer before the fully-connected layers to obtain a 2048-dimensional feature vector for each image. Convolutional neural networks have been shown to accurately capture high-level and low-level image semantics. The task of finding similar images translates into approximate nearest neighbor search on the extracted features vectors representing each image. For this, we employ hierarchical navigable small world graphs (HNSW) [9]. The resulting index is approximately 1.1 GB and a query takes less than 10ms on a modern server.

The final component of our tool is the interface, implemented in React.js and backed by a Python Flask server. A screenshot is shown in Figure 1, with a "My Little Pony" cartoon as the query image (top left). Our tool shows similar images in an "infinite scroll" progression. The user can click on any image to use it as a query. The red icon displays the number of duplicates discovered for that image, and clicking the blue icon brings up a list of all pages that contain that image with links out to the Internet Archive's Wayback machine as a rendering service for the user to examine the image in context. The tool can also select a random image in its index as the query, which supports serendipitous exploration and discovery.

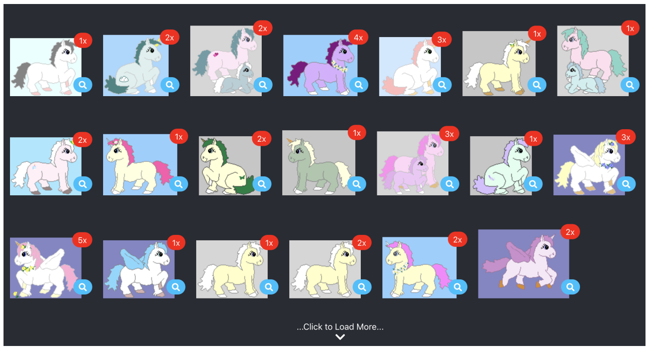

Figure 1: Screenshot of our image exploration tool.

4 ONGOING WORK

Viewed in a broader context, our work has two main goals. First, we wished to explore images in web archives and help nudge the field away from text- and link-centric analyses. Modern neural network techniques make such analyses increasingly feasible. Second, we wish to convey to the community that "obvious" renditions of existing tools (in this case, keyword-based image search), may not be the best vehicle for exploring web archives. We hope that our initial efforts inspire additional work along these lines.

Acknowledgments. This research was supported in part by the Andrew W. Mellon Foundation and the Social Sciences and Humanities Research Council of Canada.

REFERENCES

[1] T. Arnold and L. Tilton. 2019. Distant Viewing: Analyzing Large Visual Corpora. *Digital Scholarship in the Humanities* (2019).

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. 2014. Neural Codes for Image Retrieval. In *ECCV*.

[3] F. Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR*. 1800–1807.

[4] A. Cocciolo. 2015. The Rise and Fall of Text on the Web: A Quantitative Study of Web Archives. *Information Research* 20, 3 (2015).

[5] M. Costa and M. Silva. 2010. Understanding the Information Needs of Web Archive Users. In *IWAW*.

[6] H. Huurdeman, A. Ben-David, and T. Sammar. 2013. Sprint Methods for Web Archive Research. In *Web Science*.

[7] J. Lin, I. Milligan, D. Oard, N. Ruest, and K. Shilton. 2020. We Could, but Should We? Ethical Considerations for Providing Access to GeoCities and Other Historical Digital Collections. In *CHIIR*.

[8] J. Lin, I. Milligan, J. Wiebe, and A. Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *ACM JCCH* 10, 4 (2017), Article 22.

[9] Y. Malkov and D. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE PAMI* 42, 4 (2020), 824–836.

[10] I. Milligan. 2019. *History in the Age of Abundance? How the Web is Transforming Historical Research*. McGill-Queen's University Press.

[11] N. Ruest, J. Lin, I. Milligan, and S. Fritz. 2020. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. *arXiv:2001.05399* (2020).

[12] N. Ruest, I. Milligan, and J. Lin. 2019. Warclight: A Rails Engine for Web Archive Discovery. In *JCDL*.

[13] M. Wevers and T. Smits. 2019. The Visual Digital Turn: Using Neural Networks to Study Historical Images. *Digital Scholarship in the Humanities* 35, 1 (2019), 194–207.

[14] H.-W. Yang, L. Liu, I. Milligan, N. Ruest, and J. Lin. 2019. Scalable Content-Based Analysis of Images in Web Archives with TensorFlow and the Archives Unleashed Toolkit. In *JCDL*.