**A Half-Century of Illustrated Pages: ACS Lab Notes**

Stephen Krewson

We’ve reached the mid-point of my [Advanced Collaborative Support project,](https://www.hathitrust.org/hathitrust-research-center-awards-five-acs-projects) “Deriving Basic Illustration Metadata.” Right now, sitting on a supercomputer named Big Red at Indiana University, is a rather remarkable dataset: *every illustrated page from every Google- scanned volume in the HathiTrust Digital Library for the period 1800-1850.* Although the image processing pipeline we are using is not new, working at this scale is.

A working hypothesis of historical studies of illustration is that graphic elements in printed objects exhibit technical and stylistic similarity over time, but this development is *uneven*. That is to say, illustrations “change with the times” at different rates. On the one hand, the emergence of new techniques like lithography will exert a competitive pressure within the book market and prompt artisans to explore the affordances of the new method. On the other hand, reuse of woodblocks and other at-hand materials remains appealing for financial reasons. In some cases, a “vintage” style may even be desirable.

Before attempting to characterize this “unevenness” computationally, many smaller steps are necessary. The first step involves putting together a dataset of images that can plausibly represent the breadth of printed illustrations for a period in time. My project’s 50-year sample pushes up against the limits of what can be stored on disk for a project, but HTRC’s resources make possible what would be otherwise unimaginable for a graduate student. Although choosing a start year at the beginning of a century has no direct relationship to innovation in image production, 50 years is roughly the length of a [Kondratiev cycle](https://en.wikipedia.org/wiki/Kondratiev_wave). At the very least, five decades should generate plenty of real historical signal.

What do I mean by “signal”? At the end of the project, each illustration in the dataset will be encoded as a vector of numbers. This allows for calculating similarities between images. All other things being equal, we would expect similar layouts, textures, and subject matter to loosely correlate with one other in time. By building “nearest neighbor” maps for certain types of images, we can test whether this is the case and also discover moments of aesthetic anachronism, when older styles reappear. The following sections discuss our progress, making some detours to discuss problems that future researchers may face when using HathiTrust for *image* rather than text analysis.

**Choosing a Subset of Volumes**

There were 500,013qualifying volumes (as of August 2019) in HATHITRUST for the first half of the nineteenth century. To come up with this list of unique volume ids, I

filtered the latest [HathiFiles](https://www.hathitrust.org/hathifiles) by date range (1800-1850), media type, (text) and digitization agent (Google). I chose this date range because it is both relevant to my own research and because it precedes the explosion of illustrations in the second half of the nineteenth century (due in large part to half-tone printing and photography).

The vast majority of volumes for this period have been scanned as part of the Google Books project; we opted for this restriction because Google-scanned books come with some extra metadata that is useful for identifying illustrated pages. I did not perform any deduplication of volume titles.

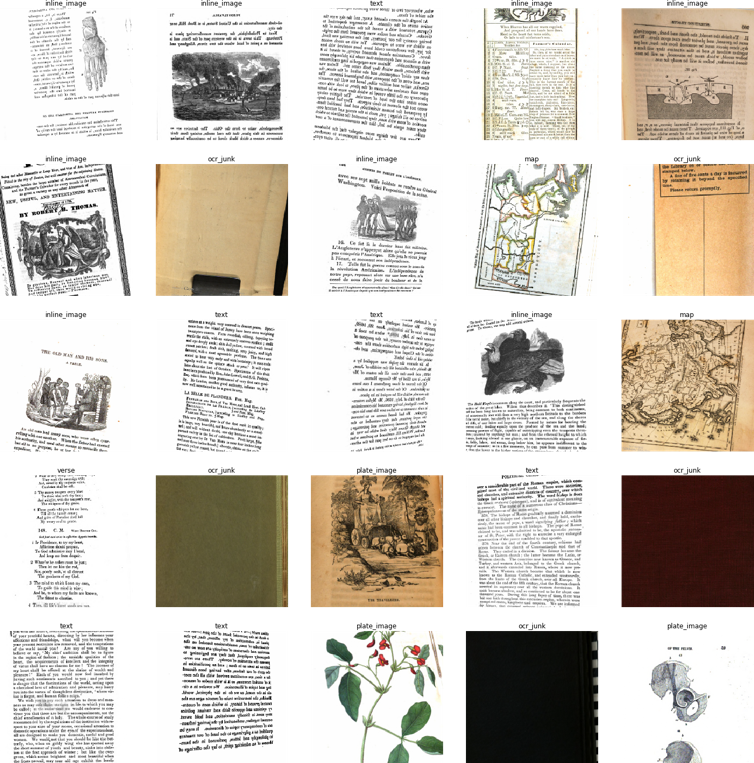
***Takeaway****:* Warning! HathiFiles are roughly a gigabyte in size–far too large to read into memory with a Pandas method like .read\_csv(). Parsing the file in chunks is the way to go (it still takes ~15 min on my i7 laptop). Writing your own script to do this is a useful exercise if you plan on working with large quantities of HathiTrust data.

Figure 1: Fastai method for viewing a data batch

**Training a Classifier**

The first stage of the project relies on a retrained convolutional neural network (CNN). The purpose of the model is to classify whether an input page image belongs to one of the desired classes: inline\_image or plate\_image. Inline images are integrated into the text of a page; plates are typically larger and occupy the entire page (in some cases they are tipped-in). Multi-page spreads and foldouts are too infrequent to merit their own class. While the process of choosing labels and labeling presents many interesting challenges, this post simply presents the parameters of the model which I retrained. The training data amounted to 12 Gb (compressed) of 19th-century book pages from Internet Archive, uploaded to a Google Cloud Cloud Platform (GCP) instance with GPU. I then used an off-the-shelf PyTorch CNN trained on ImageNet (resnet50) to reach ~94% accuracy on a held-out test set.

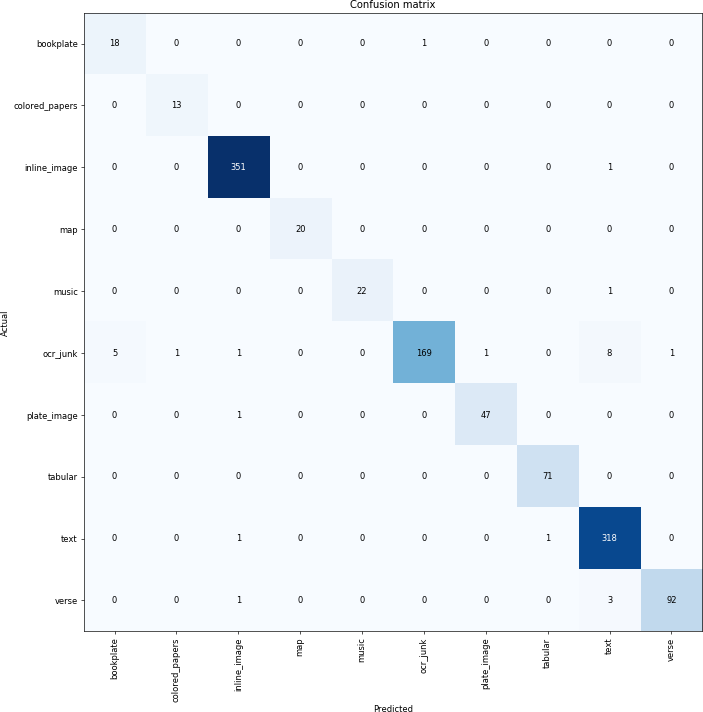
***Takeaway****:* When it comes to deep learning, try to use the collective knowledge an active community of experimenters. That’s why I chose to use [fast.ai’s](https://www.fast.ai/) PyTorch library for model training. Fast.ai provided detailed installation guides for working with GCP and the forums are invaluable for troubleshooting.

Figure 2: Confusion matrix for stage-one classifier

**Filtering out Unwanted Labels**

The trained model is saved as a serialized PyTorch object. When loaded, it provides a predict() method that returns the estimated class of the input image. Boris deserves all the credit for installing the necessary libraries (quite a headache as it turned out) and parallelizing the [code](https://github.com/htrc/ACS-krewson).

To run inference on the 500,013 volumes, we first needed to:

* Acquire volume subset from HathiTrust (in pairtree structure)
* Parse through metadata for each volume to find all pages tagged with IMAGE\_ON\_PAGE (a tag generated from OCR at digitization)
* Unzip the fold of JP2 image assets

Then HTRC ran inference with a pool of threads on high-performance compute systems at Indiana University, keeping the directory structure and deleting all volumes/assets that were not of the desired classes. When the job finished, the set of volumes was winnowed down to 183,553. This means that, according to the model, roughly 37% of early-19th century books contain one or more illustrated pages.

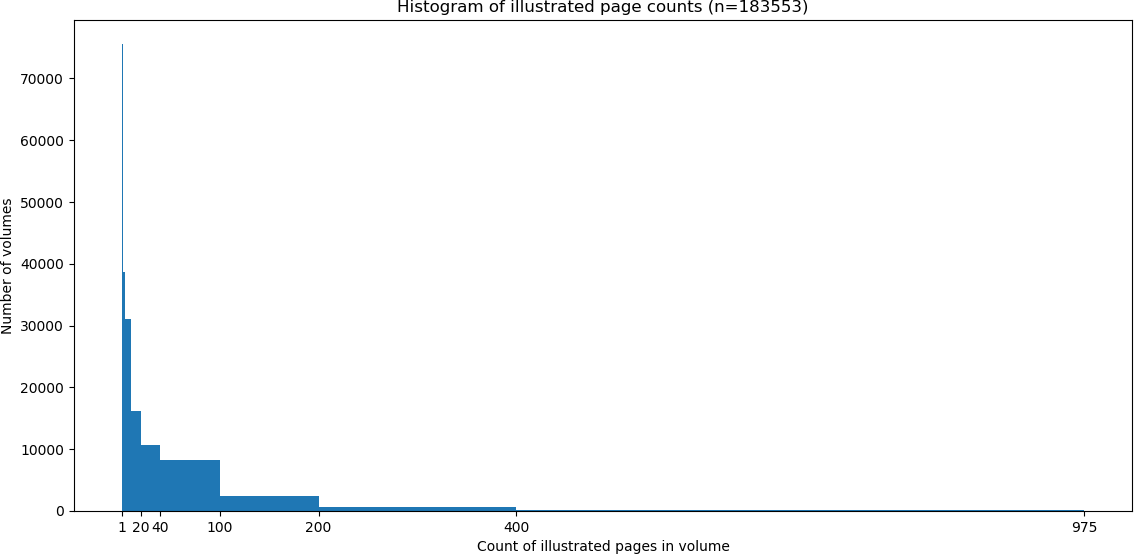
***Takeaways***: Parallelism is absolutely vital for running large jobs such as this in days instead of weeks or months. Careful thought is required regarding when to convert image formats. The JP2 format was designed for archives and is used by most HATHITRUST partners. Unfortunately, it is not supported by most machine learning libraries, including PyTorch. Thus we needed to convert to JPEG on the fly, which has been a source of endless frustration. Needs for image conversion should be considered when designing an analysis workflow.

**Initial Results**

Of this set of probably-illustrated volumes, there were 1,922,725individual pages estimated by the model to feature illustrations. That is to say, subject to survival bias and Western-centric library practices, the historical-bibliographical record for the early nineteenth century consists of about two millionprinted illustrations.

At this point, we can compute some basic statistics. The **average** number of illustrated pages for a book in this set was 10.5. The **median** number of illustrated pages was 2. The most common number (**mode**) of illustrated pages per volume was 1(e.g. only a title page or frontispiece is illustrated). The maximum count was 975. Extreme outliers are worth exploring, since they can reveal faulty assumptions about the data. In general, outliers are periodicals. If we were trying to clean up counts, we would split this into three volumes and try to (1) deduplicate and (2) standardize periodical groupings (by year, for instance). But this project is simply trying to get at the images, so the per-volume counts don’t matter all that much.

The histogram shows the distribution of illustrations: overwhelmingly, even books that are illustrated will have single-digit illustrations (most often just one or two).

****

Here’s a montage of about 500 of the images from stage one, courtesy of Damon Crockett’s ivpy Python [package:](http://damoncrockett.com/projects/ivpy.htm)



Figure 4: 533 illustrated pages extracted in phase one.

Here are the next steps of the project:

* Get bounding boxes for the images (another retrained CNN)
* “Fingerprint” the images with difference/perceptual hash or simply by using the last fully-connected layer from their representation in the stage-two neural network.
  + Define some simple experiments to assess which method is best for non-photorealistic images from historical books
* Build a lookup tree of nearest neighbors and a way to reverse image search
* Select a few clusters of stylistically similar images
  + Historicize and interpret!