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

Hsiu-Wei Yang, Linqing Liu, Ian Milligan, Nick Ruest, and Jimmy Lin

Introduction

- The lack of tools to provide scholarly access to web archiving is a big challenge for the community.
- Previous efforts focus on textual content; however, non-textual media is equally important.
- We integrate the Archives Unleashed Toolkit (https://archivesunleashed.org/aut/) with Google’s TensorFlow deep learning toolkit (https://www.tensorflow.org/).
- This combination allows scholars to directly peer into the content of images in web archives at scale.

Case study

GeoCities Collection

- The web hosting platform had seven million users and consists of 186M HTML pages. The entire web archive totals 4TB.
- Using object detection, we can find clusters of images that can suggest the existence of coherent communities.

Performance Analysis

- 2.3M images in our GeoCities archive:
  - Inference on a single image takes approximately 550ms with CPU.
  - Analyze the entire collection in a week on single high-end server.
  - This time can be greatly reduced with GPU-based inference.

Conclusions

- We exploited image analysis to counterbalance the dominance of text in digital humanities research.
- Integration of TensorFlow and AUT combines image analysis with existing capabilities — for example, enabling questions that simultaneously interrogate hyperlink structures, textual content, as well as image content.

References