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

UNIVERSITY OF WATERLOO

YORK UNIVERSITÉ UNIVERSITY

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.

Implementation

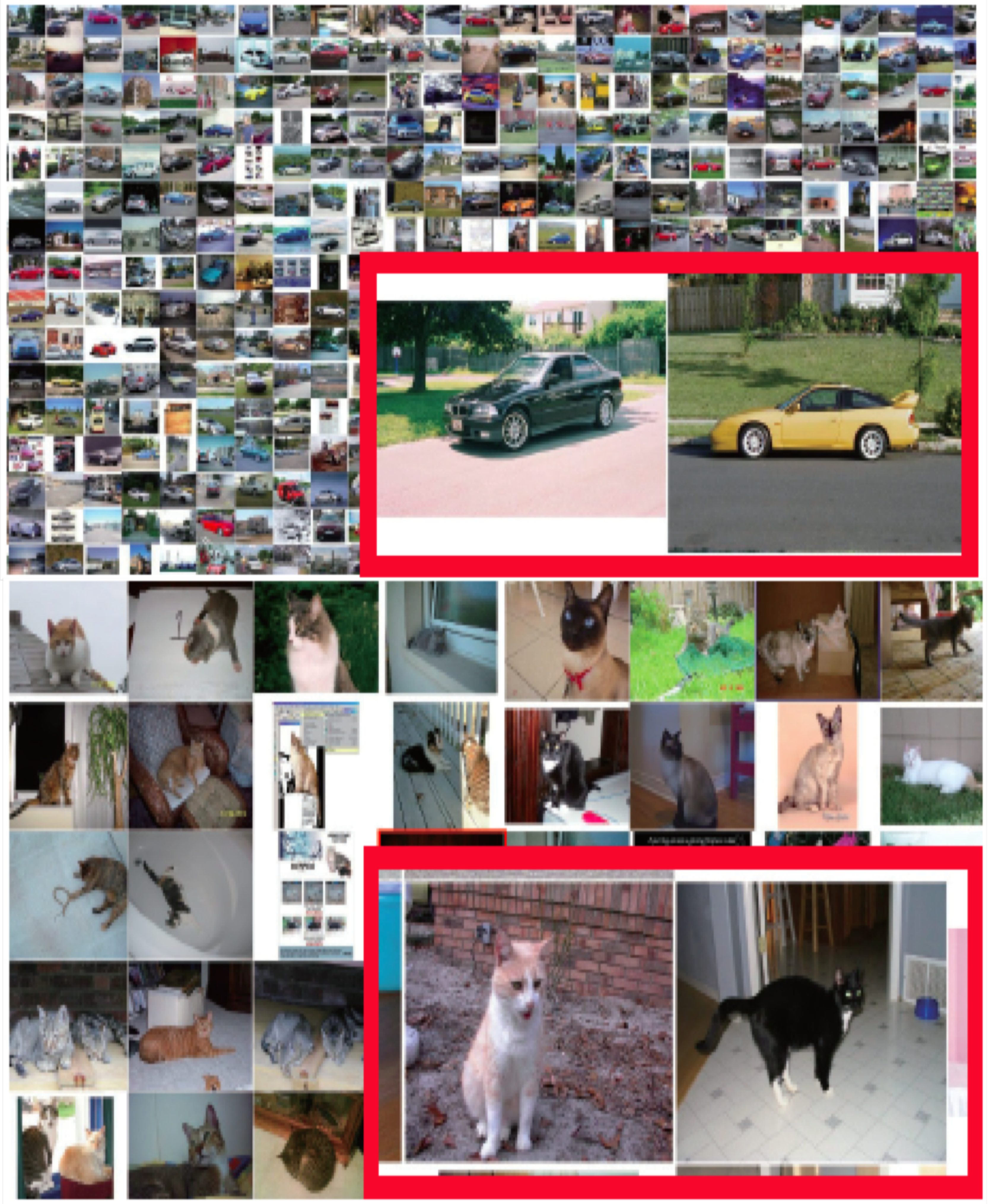
- Bridge the programming language gap between Archives Unleashed Toolkit and Tensorflow, **manipulate RDDs directly in Python**.
- Use the **Single Shot MultiBox Detector model** available in Tensorflow and broadcast it to all Spark executors to reduce inference latency.
- Model outputs are objects detected in the images and associated probabilities.

Image processing capabilities can be integrated into other Toolkit analyses.

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.
- **Check it out at https://ruebot.net/geocities-jcdl2019/**

References

[1] Lin, Jimmy, et al. "Warcbase: Scalable analytics infrastructure for exploring web archives." Journal on Computing and Cultural Heritage (JOCCH) 10.4 (2017): 22.

[2] I. Milligan. 2019. GeoCities. In SAGE Handbook of Web History, Niels Brügger and Ian Milligan (Eds.). SAGE Publications, London.

[3] Liu, Wei, et al. "Ssd: Single shot multibox detector." In ECCV 2016.