An Approach for Automatic and Large Scale Image Forensics

Thamme Gowda*, Kyle Hundman #, and Chris Mattmann **

Presenter: Paul Ramirez #

* Computer Science Department, University of Southern California, Los Angeles, CA, USA

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA, USA
OVERVIEW

• Abstract
• Motivation
• Data
• Image Recognition
• Inception Net
• Integration
• Evaluation
• Conclusion
ABSTRACT

• Applications of deep learning-based image recognition in the DARPA Memex program
• Integration of Tensorflow with Apache Tika for automatic image forensics
• Evaluation of model performance on weapons dataset
MOTIVATION

DARPA Memex:
• Monitor online weapons sale in the United States
• Goal 1: Retrieve ads and relevant multimedia such as images, videos
• Goal 2: Forensics
  • Classify illegal weapons
  • Sale trends
• Goal 3: Discoverable / Searchable
DATA COLLECTION

- Used web crawlers specialized for retrieving data
  - Crawlers that can login to web sites and run javascript in pages
  - Crawlers that can work with Onion protocol
  - Example: Apache Nutch, Sparkler, Scrapy, ... by various teams
- Large repository of web pages and multimedia documents
  - 1.4 M images from weapons domain
IMAGE RECOGNITION TASK

• Image Recognition: Detect real word entities in the digital images
• ImageNet dataset:
  • Large visual dataset of annotated images
• ImageNet Large Scale Visual Recognition Challenge (ILSVRC)
  • Annual competition organized by Stanford and Princeton Universities
  • Challenge: How accurately can your model identify 1000 classes
  • From 2010 to Now
  • Since 2012, Deep ConvNets ruled the competition
  • Goto place to see state-of-the-art models for image recognition
INCEPTION NET

• Developed by Google Research Team
  • Sergey et al, 2015 - Originally GoogleNet, winner of ILSVRC 2014
  • Code named Inception, multiple version V1, V2, V3, V4,..
• Google open sourced Tensorflow with Inception-V3 and its model trained on ImageNet dataset
• Inception-V3 is optimized to run with less memory and fewer CPU cycles (like Android devices)
• We have used Inception-V3 for our forensics
SOFTWARE STACK

• Apache Tika - universal parser for parsing files over a thousand file types
  • Primarily written in Java; available for free via Apache License
  • Meta data analysis
  • Semantic analysis - detect names of people, locations etc in text
  • And more - OCR in images
  • One of the key technology for content analysis in DARPA Memex
  • Had been useful for others too - heard of Panama Papers?

• Tensorflow
  • Written in C++ with Python bindings; available free via Apache License
  • Developed by google and now one of the popular deep learning frameworks
INTEGRATION METHODS

• Challenge:
  • Make use of C++/Python code from a Java Client
• Techniques
  • Command Line Invocation (CLI)
  • Java Native Interface (JNI)
  • gRPC Remote Procedure Call (gRPC)
  • REpresentation State Transfer (REST) API
• REST API integration was the best among the above four
RESULTS

Labeled the 1.4 million images in Memex Weapons Dataset

1000 target classes in training data (ImageNet)
HANDLING THE SCALE

• 1.4 million images in the dataset
• REST integration took 36 hours to run on 32 Core CPUs, no GPUs used
• TensorFlow automatically parallelized the load on all CPU cores in a single node
• Wiki [https://wiki.apache.org/tika/TikaAndVision](https://wiki.apache.org/tika/TikaAndVision)

• **Recent work:** We have hadoop/spark distributable framework powered by Deeplearning4j
  • [https://github.com/thammegowda/tika-dl4j-spark-imgrec](https://github.com/thammegowda/tika-dl4j-spark-imgrec)
Our evaluation dataset:
- Consists of gun images
- Law enforcement officers manually labelled them

Observations:
- Some Rifles mislabeled based on surrounding objects - small size
- Top - 5 measure is a reasonable measure
CONCLUSION

• We have made image recognition easy for Apache Tika users
• We have tested that Inception-V3 model was successful in detecting weapon images
• Image labels helped to build a better web page classifier for Memex

ACKNOWLEDGEMENT:
This effort was supported in part by JPL, managed by the California Institute of Technology on behalf of NASA, and additionally in part by the DARPA Memex/XDATA/D3M programs and NSF award numbers ICER-1639753, PLR-1348450 and PLR-144562 funded a portion of the work
THANKS