WARC-DL: Scalable Web Archive Processing for Deep Learning

Niklas Deckers, Martin Potthast

Leipzig University
Webis Group
webis.de

OSSYM 2022 – 4th International Open Search Symposium, October 10, 2022
Given a learning task and ground truth within WARC files, train a model. Only a fraction of the records within the WARC files are ground truth.

Goal: Training at web scale (billions of WARC files)
Web Archive Processing

Mining

- Given a mining task and a trained (classification) model, collect relevant data. Only a fraction of the records within the WARC files are relevant.

- Goal: Mining at web scale (billions of WARC files)
Given a mining task and a trained (classification) model, collect relevant data. Only a fraction of the records within the WARC files are relevant.

Goal: Mining at web scale (billions of WARC files)

Observations:
- Mining / filtering WARC files is “embarrassingly parallel”.
- Decompressing WARC files, and processing WARC records are CPU bound.
- The preprocessing step results in a variable data flow.
- Training of neural networks is GPU bound and presumes constant data flow.
- WARC storage, parallel processing, and GPU bound processing are on separate clusters.
## Webis Data Center (Digital Bauhaus Lab)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td><strong>Disk [PB]</strong></td>
<td>0.2</td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Cores</strong></td>
<td>176</td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td><strong>RAM [TB]</strong></td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≅ 3.2 TFLOPs</td>
<td>≅ 67.4 TFLOPs</td>
<td>≅ 8 PFLOPs</td>
<td>≅ 119.8 TFLOPs</td>
</tr>
</tbody>
</table>

**Typical research:**

- **α-Web.** Teaching, Staging environment
- **β-Web.** Web mining (map reduce), CPU parallelization
- **γ-Web.** Machine learning (embedding, deep learning), Language modeling
- **δ-Web.** Web archive storage (10 PB from Internet Archive and Common Crawl)
- **ε-Web.** Search index construction, Argument search
1. PySpark distributes WARC streams among workers
2. FastWARC decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization
Web Archive Processing
WARC-DL: Pipeline for Processing at Petabyte Scale

1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization
3. Pickled record streams
1. PySpark distributes WARCs among workers
2. **FastWARC** decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization
3. Pickled record streams
4. Conversion to TensorFlow datasets and source interleaving
5. Inference: Batched processing by a Keras model
   and second filtering based on classification results
1. PySpark distributes WARC
distributes WARC among workers

2. FastWARC decompresses and iterates records
   CPU-bound filtering, feature extraction, tokenization

3. Pickled record streams

4. Conversion to TensorFlow datasets and source interleaving

5. Inference: Batched processing by a Keras model
   and second filtering based on classification results

6. Optional filtering (e.g., deduplication) and model training
Application: Building Large-Scale Multimodal Datasets
For Training Generative Text-To-Image Models

- CompVis group created the Latent Diffusion model
- LAION created a dataset of text-image pairs
  Consists of image urls and img alt attribute texts from Common Crawl
- Stability AI finetuned Latent Diffusion on this dataset to create Stable Diffusion

Image generated by Stable Diffusion with the prompt
“award-winning cake shaped like the Swiss Alps”
Application: Building Large-Scale Multimodal Datasets
For Training Generative Text-To-Image Models

- CompVis group created the Latent Diffusion model
- LAION created a dataset of text-image pairs
  Consists of image urls and img alt attribute texts from Common Crawl
- Stability AI finetuned Latent Diffusion on this dataset to create Stable Diffusion
- Next target together with LAION: Building a better multimodal dataset
- Obtaining such a dataset requires preprocessing, rule-based and DL-based filtering (e.g., NSFW filtering)
  Using the WARC-DL pipeline allows quick deployment on existing infrastructure
- Include text, images, videos and audio
- Extract more context from around the media links
  Will enable text-to-image models to work with more complex prompts
Conclusion

WARC-DL can be used for petascale web archive processing:

- Training and applying domain-specific models for web mining
- Dataset extraction
- (Multimodal) Search engines
  Will be applied in the upcoming Open Web Search project
Conclusion

WARC-DL can be used for petascale web archive processing:

- Training and applying domain-specific models for web mining
- Dataset extraction
- (Multimodal) Search engines
  Will be applied in the upcoming Open Web Search project

Thank you!