DeepJetCore

Swapneel Mehta, Mauro Verzetti, Jan Kieseler, Markus Stoye, Maurizio Pierini
CMS Experiment, EP-CMG-PS
CERN
Machine Learning

1. Comprehensive libraries
2. Fantastic documentation
3. Interactive Tutorials
4. Developer Community Support
Why build a library designed for high-energy physics?
Computer Scientists don’t always understand requirements for particle physics...

Why is CERN trying to open a portal to allow the dark entities access to our dimension?
Physicists don’t always write great code...

```java
public Date getNextDay() {
    try {
        Thread.sleep(TimeUnit.DAYS.toMillis(1));
        return new Date(); //success
    } catch (InterruptedException e) {
        e.printStackTrace();
        return null; //failure
    }
}
```
Best of Both Worlds

1. Implement fast, efficient machine learning algorithms for physics
2. Provide high-level functions/wrappers for low-level tasks
3. Handle common bottlenecks - esp. memory-related issues
4. Create an extensible, easy-to-use framework
What does this library do?
Features of DeepJet

- Data Conversion
- Model Training
- Prediction
- Model Evaluation
- File-by-File
- Avoids memory threshold crossed (EOS)
- Handles user-defined data structures
- Preprocessing support
- Parallelized operation
- Keras-wrapped Tensorflow backend
- Additional callbacks
- Monitor validity of tokens
- Bookkeeping support
- Create compatible prediction data structures
- Support for Plots
- Export of models and data structures
Yeah, but why should I use it?
- Modularised code, easy to understand
- Templates for quick-start
- Step-by-step documentation
- Elaborate examples and use-cases
• C++ Extensions improve efficiency for Python
• Anaconda support
• Available as a Python Package
• Docker support
Interesting! Tell me more about this library
## DeepJetCore

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepJetCore</td>
<td>override build_ext for custom make</td>
</tr>
<tr>
<td>bin</td>
<td>modified README, added Apache License</td>
</tr>
<tr>
<td>.gitignore</td>
<td>modified cmdclass adding custom build in setup</td>
</tr>
<tr>
<td>CHANGELOG.rst</td>
<td>breaking changes for setup and compilation of scripts as extensions</td>
</tr>
<tr>
<td>MANIFEST.in</td>
<td>modified MANIFEST to add swig wrappers in package</td>
</tr>
<tr>
<td>README.md</td>
<td>restructuring and modifications to setup</td>
</tr>
<tr>
<td>README.rst</td>
<td>modified README, added Apache License</td>
</tr>
<tr>
<td>environment.yml</td>
<td>modified setup for manual build</td>
</tr>
<tr>
<td>setup.cfg</td>
<td>modified setup for manual build</td>
</tr>
<tr>
<td>setup.py</td>
<td>override build_ext for custom make</td>
</tr>
</tbody>
</table>
little tool to add prediction labels to a data collection - only need...

OSX setup

OSX setup

fix

more checks and fixes. fixed other global treename per traindata

better handling of prediction labels, more options, better bookkeeping...

updates for domain adaptation

better handling of prediction labels, more options, better bookkeeping...

adding license

not working restructuring commit

Update README.md

better handling of prediction labels, more options, better bookkeeping...

not working restructuring commit

not working restructuring commit

more checks and fixes. fixed other global treename per traindata

more checks and fixes. fixed other global treename per traindata

not working restructuring commit

not working restructuring commit
<table>
<thead>
<tr>
<th>Directory</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>fixed makefile</td>
</tr>
<tr>
<td>Train</td>
<td>better label bookkeeping - connected to Core, and bugfixes for dom ad...</td>
</tr>
<tr>
<td>independence/example</td>
<td>decorrelation and independence</td>
</tr>
<tr>
<td>modules</td>
<td>better label bookkeeping - connected to Core, and bugfixes for dom ad...</td>
</tr>
<tr>
<td>scripts</td>
<td>removed core part</td>
</tr>
<tr>
<td>.gitignore</td>
<td>yet another update</td>
</tr>
<tr>
<td>LICENSE.txt</td>
<td>added license</td>
</tr>
<tr>
<td>README.md</td>
<td>Update README.md</td>
</tr>
<tr>
<td>env.sh</td>
<td>one-size-fit-all environment</td>
</tr>
<tr>
<td>gpu_env.sh</td>
<td>more restructuring</td>
</tr>
<tr>
<td>lxplus_env.sh</td>
<td>more restructuring</td>
</tr>
</tbody>
</table>
• Install not entirely automated yet

• Platform supported: CentOS 7

• Breaking changes in recent releases of Tensorflow
Conclusion

- Easy-to-use Framework
- Faster conversion and training
- Diverse use-cases
- Scalable to large datasets
Thank You!