Open source tools for machine learning models and data sets versioning

Dmitry Petrov iterative.AI
Dmitry Petrov
PhD in Computer Science
Twitter: @FullStackML

Co-Founder & CEO > Iterative.AI > San Francisco, USA

ex-Data Scientist > Microsoft (BingAds) > Seattle, USA

ex-Head of Lab > St. Petersburg Electrotechnical University > Russia
Do we need new tools for ML?

- MLFlow
- Git-LFS
- DVC
- Conclusion
> Do we need new tools for ML?
> MLFlow
> Git-LFS
> DVC
> Conclusion
1/The rise of Software Engineering required inventing processes like version control, code review, agile, to help teams work effectively. The rise of AI & Machine Learning Engineering is now requiring new processes, like how we split train/dev/test, model zoos, etc.
PROBLEM 1: ML IS SLOW

MY MODEL'S TRAINING
"MY CODE'S Compiling."

Solution: custom ML PIPELINES

© xkcd
PROBLEM 2: ML IS METRICS DRIVEN

>> EXPERIMENT = CODE + OUTPUTS

Outputs include metrics and graphs AUC, etc.

Solution: metrics tracking
**Problem 3: Mess with Data Artifacts**

>> EXPERIMENT = CODE + OUTPUTS + DATASET

Source code, Datasets, ML models

**Solution:** connect data to code
<table>
<thead>
<tr>
<th>SUMMARY OF DIFFERENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software engineering</strong></td>
</tr>
<tr>
<td>Source code version control</td>
</tr>
<tr>
<td>Code review</td>
</tr>
<tr>
<td>Agile methodology</td>
</tr>
<tr>
<td>-_(ツ)_/\ -</td>
</tr>
</tbody>
</table>
Do we need new tools for ML?

- MLFlow
- Git-LFS
- DVC
- Conclusion
Platform for the machine learning lifecycle

> Tracking
> Project
> Models

$ pip install mlflow
from mlflow import log_metric, log_param, log_artifact

log_param("lr", 0.03)
log_metric("loss", curr_loss)
log_artifact("loss")

$ mlflow ui
## MLFLOW TRACKING UI

<table>
<thead>
<tr>
<th>Date</th>
<th>User</th>
<th>Source</th>
<th>Version</th>
<th>Parameters</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018-06-04 23:00:10</td>
<td>mlflow</td>
<td>train.py</td>
<td>05e956</td>
<td>1 1</td>
<td>0.649 0.04 0.862</td>
</tr>
<tr>
<td>2018-06-04 23:00:10</td>
<td>mlflow</td>
<td>train.py</td>
<td>05e956</td>
<td>1 0.5</td>
<td>0.648 0.046 0.859</td>
</tr>
<tr>
<td>2018-06-04 23:00:10</td>
<td>mlflow</td>
<td>train.py</td>
<td>05e956</td>
<td>1 0.2</td>
<td>0.628 0.125 0.823</td>
</tr>
<tr>
<td>2018-06-04 23:00:09</td>
<td>mlflow</td>
<td>train.py</td>
<td>05e956</td>
<td>1 0</td>
<td>0.619 0.176 0.799</td>
</tr>
<tr>
<td>Feature</td>
<td>Result</td>
<td>Comment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------</td>
<td>---------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Versioning ML models</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Versioning datasets</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Versioning ML pipelines</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connecting data and code</td>
<td>-/+</td>
<td>Manual only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tracking metrics</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visualize metrics</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
> Do we need new tools for ML?
> MLFlow
> Git-LFS Git Large File Storage
> DVC
> Conclusion
> Install

$ brew install git-lfs
$ git lfs install

> Specify data-files type in a Git repository

$ git lfs track '*.p'
$ git add .gitattributes
$ python mytrain.py  # your code generates mymodel.p
$ git add mytrain.py mymodel.p
$ git commit -m 'Decay was added'
$ git push

Uploading LFS objects: 100% (1/1),
56 MB  | 3.2 MB/s, done
$ git clone https://github.com/dmpetrov/my-lfs-repo
$ cd my-lfs-repo
$ du -sh mymodel.p  # data file does not contain data yet
  4.0K  mymodel.p
$ git pull

Downloading LFS objects: 75% (3/4),
44 MB | 4.5 MB/s
| GIT-LFS PROS/CONS |

> **PROS**
> Simple, like Git

> **CONS**
> Limited by data size  <2Gb, <500Mb even better
> Not every Git server supports Git-LFS
> No ML\Data Science specific
<table>
<thead>
<tr>
<th>Feature</th>
<th>Result</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Versioning ML models</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Versioning datasets</td>
<td>-/+</td>
<td>Limited by size</td>
</tr>
<tr>
<td>Versioning ML pipelines</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Connecting data and code</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Tracking metrics</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Visualize metrics</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Do we need new tools for ML?

- MLFlow
- Git-LFS
- DVC Data Version Control
- Conclusion
Website: http://DVC.org

> Install

```
$ pip install dvc
$ dvc init
```

> Git-like tool no infrastructure is required
DVC ADD DATA FILES

> Push data to storage

```bash
$ dvc add data.xml
$ dvc push
```

> Push meta information to Git server

```bash
$ git add .gitignore data.xml.dvc
$ git commit -m "add source data to DVC"
$ git push
```
$ git clone https://github.com/dmpetrov/my-dvc-repo
$ cd my-dvc-repo
$ dvc pull
...

$ du -sh data.xml
7G  data.xml
```bash
$ git clone https://github.com/dmpetrov/my-dvc-repo
$ cd my-dvc-repo
$ dvc pull train.dvc
...

$ du -sh cnn_model.p
  54M  cnn_model.p
```
Checkout data

```text
$ git checkout vgg16_exp2
$ dvc checkout
```
Copy 50G directory with millions of images ~10 min

What about DVC?

```
$ git checkout image_update_20190310
$ time dvc checkout

real 0m12.958s
user 0m11.567s
sys 0m1.725s
```
$ dvc add data/data.xml
$ dvc run -d src/prepare.py -d data/data.xml -o data/prepared \
    python src/prepare.py data/data.xml
$ dvc run -d src/featurization.py -d data/prepared -o \
    data/features \
    python src/featurization.py data/prepared data/features
$ dvc run -d src/train.py -d data/features -o model.pkl \
    python src/train.py data/features model.pkl
$ dvc pipeline show --ascii train.dvc --commands

+-------------------------------------+
| python src/prepare.py data/data.xml |
+-------------------------------------+

*  
*  
*  

+---------------------------------------------------------+
| python src/featurization.py data/prepared data/features |
+---------------------------------------------------------+

*  
*  
*  

+---------------------------------------------+
| python src/train.py data/features model.pkl |
+---------------------------------------------+
DVC PIPELINES: REPRODUCIBILITY

> Reproduce your project
   $ dvc repro

> Reproduce
   $ dvc repro train.dvc

> Version DVC pipeline
   $ git add train.dvc
   $ git commit -m 'Reproduce with dataset update 2019-05-02'
<table>
<thead>
<tr>
<th>Feature</th>
<th>Result</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Versioning ML models</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Versioning datasets</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Versioning ML pipelines</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Connecting data and code</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Tracking metrics</td>
<td>−/+</td>
<td>Final metrics only</td>
</tr>
<tr>
<td>Visualize metrics</td>
<td>−</td>
<td></td>
</tr>
</tbody>
</table>
> Do we need new tools for ML?
> Git-LFS
> MLFlow
> DVC

> Conclusion
<table>
<thead>
<tr>
<th>Feature</th>
<th>MLFLOW</th>
<th>Git-LFS</th>
<th>DVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Versioning ML models</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Versioning datasets</td>
<td>-</td>
<td>-/+</td>
<td>+</td>
</tr>
<tr>
<td>Versioning ML pipelines</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Connecting data and code</td>
<td>-/+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Tracking metrics</td>
<td>+</td>
<td>-</td>
<td>-/+</td>
</tr>
<tr>
<td>Visualize metrics</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Data science as different from software as software was different from hardware

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Software</th>
<th>DS/ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterfall</td>
<td>Agile</td>
<td>_(ツ)_/\</td>
</tr>
</tbody>
</table>
Think about processes
Try new ML tools
Share your feedback
THANK YOU

> Questions

Twitter @FullStackML
Email dmitry@iterative.ai

> Actions

Visit dvc.org
Star github.com/iterative/dvc
Appendix
2/I'm also seeing many AI teams use new processes that haven't been formalized or named yet, ranging from how we write product requirement docs to how we version data and ML pipelines. This is an exciting time for developing these ideas!