Open source tools for ML experiments management

Dmitry Petrov & Ruslan Kuprieieiev

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
| AGENDA |

> Why ML is special?
> MLFlow
> Git-LFS
> DVC
> Conclusion
01

> Why ML is special?
> 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.
DIFFERENCE 1: HYPER PARAMETERS

Hyperparameters evolve faster than code.
Hyperparameters evolve faster than code.

<table>
<thead>
<tr>
<th>Date</th>
<th>Alpha</th>
<th>L1_ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-04-03 1:00 PM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2019-04-03 4:00 PM</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2019-04-04 9:00 AM</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2019-04-04 11:00 AM</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
DIFFERENCE 2: ML IS METRICS DRIVEN

This is how we navigate.
**DIFFERENCE 2: ML IS METRICS DRIVEN**

**Solution:** metrics tracking

<table>
<thead>
<tr>
<th>Date</th>
<th>Alpha</th>
<th>L1_ratio</th>
<th>mea</th>
<th>r2</th>
<th>mse</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-04-03 1:00 PM</td>
<td>1</td>
<td>1</td>
<td>0.649</td>
<td>0.04</td>
<td>0.862</td>
</tr>
<tr>
<td>2019-04-03 4:00 PM</td>
<td>1</td>
<td>0.5</td>
<td>0.648</td>
<td>0.046</td>
<td>0.859</td>
</tr>
<tr>
<td>2019-04-04 9:00 AM</td>
<td>1</td>
<td>0.2</td>
<td>0.628</td>
<td>0.125</td>
<td>0.823</td>
</tr>
<tr>
<td>2019-04-04 11:00 AM</td>
<td>1</td>
<td>0</td>
<td>0.619</td>
<td>0.176</td>
<td>0.799</td>
</tr>
</tbody>
</table>
DIFFERENCE 3: ML MODELS CENTRIC

SLOW ML MODELS TRAINING TIME
DIFFERENCE 3: ML MODELS CENTRIC

SLOW ML MODELS TRAINING TIME

Solution: copying and versioning ML models

<table>
<thead>
<tr>
<th>Date</th>
<th>Model</th>
<th>Alpha</th>
<th>L1_ratio</th>
<th>mea</th>
<th>r2</th>
<th>mse</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-04-03 1:00 PM</td>
<td>model_v7.p</td>
<td>1</td>
<td>1</td>
<td>0.649</td>
<td>0.04</td>
<td>0.862</td>
</tr>
<tr>
<td>2019-04-03 4:00 PM</td>
<td>model_v7_l1-05.p</td>
<td>1</td>
<td>0.5</td>
<td>0.648</td>
<td>0.046</td>
<td>0.859</td>
</tr>
<tr>
<td>2019-04-04 9:00 AM</td>
<td>model_v7_l1-02_d3.p</td>
<td>1</td>
<td>0.2</td>
<td>0.628</td>
<td>0.125</td>
<td>0.823</td>
</tr>
<tr>
<td>2019-04-04 11:00 AM</td>
<td>model_v7_l1-zero_d3.p</td>
<td>1</td>
<td>0</td>
<td>0.619</td>
<td>0.176</td>
<td>0.799</td>
</tr>
</tbody>
</table>
DIFFERENCE 4: DATASETS

Which version I used last week?
DIFFERENCE 4: DATASETS

> Size is usually large > Gb
> Moving datasets around
> Datasets evolve. So, versioning is needed
DIFERENCE 4: DATASETS

Solution: copying datasets & versioning the copies

<table>
<thead>
<tr>
<th>Date</th>
<th>Dataset</th>
<th>Model</th>
<th>Alpha</th>
<th>L1_ratio</th>
<th>mea</th>
<th>r2</th>
<th>mse</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-04-03 1:00 PM</td>
<td>data_v2</td>
<td>model_v7.p</td>
<td>1</td>
<td>1</td>
<td>0.649</td>
<td>0.04</td>
<td>0.862</td>
</tr>
<tr>
<td>2019-04-03 4:00 PM</td>
<td>data_v2</td>
<td>model_v7_l1-05.p</td>
<td>1</td>
<td>0.5</td>
<td>0.648</td>
<td>0.046</td>
<td>0.859</td>
</tr>
<tr>
<td>2019-04-04 9:00 AM</td>
<td>data_v2_upd_May</td>
<td>model_v7_l1-02_d3.p</td>
<td>1</td>
<td>0.2</td>
<td>0.628</td>
<td>0.125</td>
<td>0.823</td>
</tr>
<tr>
<td>2019-04-04 11:00 AM</td>
<td>data_v2_upd_May</td>
<td>model_v7_l1-zero_d3.p</td>
<td>1</td>
<td>0</td>
<td>0.619</td>
<td>0.176</td>
<td>0.799</td>
</tr>
</tbody>
</table>
Use version control: Git and others.

Problem: align code, data, metrics versions.
Use version control: Git and others.

Problem: align code, data, metrics versions.
THE SPREADSHEET ISSUE: LACK OF TRUST

People are involved in the loop:
- You need to force people to do this
- Easy to make a mistake

<table>
<thead>
<tr>
<th>Date</th>
<th>Dataset</th>
<th>Model</th>
<th>Alpha</th>
<th>L1_ratio</th>
<th>mea</th>
<th>r2</th>
<th>mse</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-04-03 1:00 PM</td>
<td>data_v2</td>
<td>model_v1</td>
<td>1</td>
<td>1</td>
<td>0.649</td>
<td>0.04</td>
<td>0.862</td>
</tr>
<tr>
<td>2019-04-03 4:00 PM</td>
<td>data_v2</td>
<td>model_v7_l1-05.p</td>
<td>1</td>
<td>0.5</td>
<td>0.648</td>
<td>0.046</td>
<td>0.859</td>
</tr>
<tr>
<td>2019-04-04 9:00 AM</td>
<td>data_v2_upd_May</td>
<td>model_v7_l1-02_d3.p</td>
<td>1</td>
<td>0.2</td>
<td>0.628</td>
<td>0.125</td>
<td>0.823</td>
</tr>
<tr>
<td>2019-04-04 11:00 AM</td>
<td>data_v2_upd_May</td>
<td>model_v7_l1-zero_d3.p</td>
<td>1</td>
<td>0</td>
<td>0.619</td>
<td>0.176</td>
<td>0.799</td>
</tr>
</tbody>
</table>
WHAT IS ML EXPERIMENT?

Code

Data

Hyper params

ML model

Metrics

TRAIN
Why ML is special?

MLFlow

Git-LFS

DVC

Conclusion
Platform for the machine learning lifecycle

- Tracking
- Project
- Models

$ pip install mlflow
MLFLOW PRINCIPLES

> HYPER-PARAMETERS FOCUSED

  > Hyperparameter tuning

> CAPTURE OUTPUTS

  > Metrics & ML models

> INTERNAL DATABASE

  > Copy of the captured params, metrics & models
from mlflow import log_metric, log_param, log_artifact
log_param("lr", 0.03)
log_metric("loss", curr_loss)
log_artifact("model.p")

$ 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>alpha: 1, l1_ratio: 1</td>
<td>mae: 0.649, r2: 0.04, rmse: 0.862</td>
</tr>
<tr>
<td>2018-06-04 23:00:10</td>
<td>mlflow</td>
<td>train.py</td>
<td>05e956</td>
<td>alpha: 1, l1_ratio: 0.5</td>
<td>mae: 0.648, r2: 0.046, rmse: 0.859</td>
</tr>
<tr>
<td>2018-06-04 23:00:10</td>
<td>mlflow</td>
<td>train.py</td>
<td>05e956</td>
<td>alpha: 1, l1_ratio: 0.2</td>
<td>mae: 0.628, r2: 0.125, rmse: 0.823</td>
</tr>
<tr>
<td>2018-06-04 23:00:09</td>
<td>mlflow</td>
<td>train.py</td>
<td>05e956</td>
<td>alpha: 1, l1_ratio: 0</td>
<td>mae: 0.619, r2: 0.176, rmse: 0.799</td>
</tr>
</tbody>
</table>

*from: mlflow.org*
> Why ML is special?
> MLFlow
> Git-LFS  Git Large File Storage
> DVC
> Conclusion
**GIT-LFS INTRO**

> Install

```bash
$ brew install git-lfs
$ git lfs install
```

> Specify data-files type in a Git repository

```bash
$ 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
Why ML is special?
MLFlow
Git-LFS
DVC Data Version Control
Conclusion
ML PROJECT WITH NO DATA VERSIONING

from: https://www.theverge.com/2016/5/5/11592622/this-is-fine-meme-comic
DVC INTRO

Website: http://DVC.org

> Install through PIP of deb\windows packages

$ pip install dvc  
$ dvc init  
$ dvc remote add -d myremote s3://bucket/path

> Programming language agnostic  
  > Python, R, Shell, Scala/Spark
Remote

Code
Github, Gitlab, any Git Server

Data
S3, Azure, Google Cloud, SSH

Local

code

model.pkl.dvc
1KB

model.pkl
500MB

git push

git pull

dvc push

dvc pull

.git

model.pkl.dvc

model.pkl

---

The diagram illustrates a workflow for managing code and data using Git and DVC (Deep Value Chain) for version control. The code is stored on remote servers like Github or Gitlab, while the data can be found on cloud storage such as S3, Azure, Google Cloud, or SSH. The `git push` command is used to upload changes to the remote code repository, while `git pull` is used to download updates from the remote repository. DVC uses `dvc push` to upload datasets and models, and `dvc pull` to download them. The `model.pkl.dvc` file is a DVC configuration file that tracks the data and model files.
DVC PRINCIPLES

> LARGE DATASETS AND ML MODELS
  > 100Gb+ datasets, 1M images directory
  > Cloud storages support: S3, Azure, GCS, SSH

> GIT AS A FOUNDATION
  > Best engineering practices
  > ML experiments compatible with DevOps tools

> ML SPECIFIC
  > Metrics and ML pipelines
DVC - ADD DATAFILE

> Push data to storage

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

> Push meta information to Git server

```
$ git add .gitignore data.xml.dvc
$ git commit -m "add source data to DVC"
$ git push
```
Push data to storage

```
$ dvc add images/
$ dvc push
```

Meta info to Git server

```
$ git add .gitignore images.dvc
$ git commit -m "add dir with images"
$ git push
```
$ git clone https://github.com/dmpetrov/my-dvc-repo
$ cd my-dvc-repo
$ dvc pull
...

$ du -sh data.xml
7G  data.xml
$ 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
$ dvc get https://github.com/dmpetrov/my-dvc-repo \
cnn_model.p
$ du -sh cnn_model.p
54M cnn_model.p
$ 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

# modify data or code and reproduce the result
$ dvc repro
Checkout data

$ git checkout vgg16_exp2
$ dvc checkout
> Best practices for ML
> Git-LFS
> MLFlow
> DVC
> Conclusion
EXPERIMENTATION TOOLS

- GitLFS
- DVC
- MLFlow

TRAIN

Code → Data → Hyper params → ML model → Metrics

EXPERIMENTATION TOOLS
Data science as different from software as software was different from hardware

Nick Elprin, Domino Data Lab

<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>
HOW TO DESIGN OUR FUTURE
Analyse the ML processes
HOW TO DESIGN OUR FUTURE

> Analyse the ML processes

> Try / develop new ML tools
How to design our future

> Analyse the ML processes
> Try / develop new ML tools
> Share your knowledge
Analyse the ML processes
Try / develop new ML tools
Share your knowledge
Support open source
THANK YOU

Questions
Twitter @FullStackML
Email dmitry@iterative.ai

Actions
Visit dvc.org
Star github.com/iterative/dvc