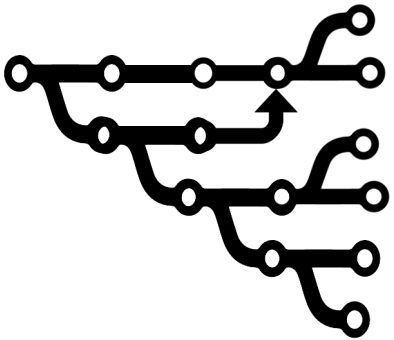
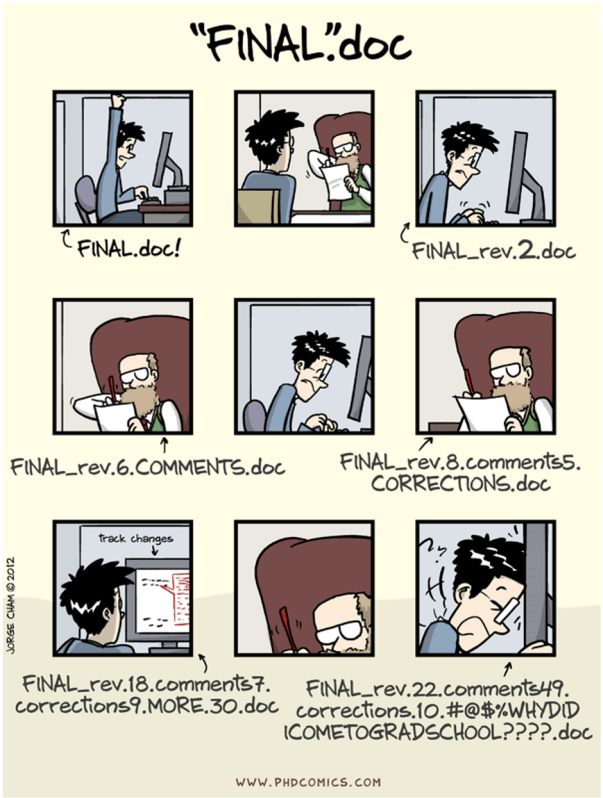
Versioning ML Models & Data in Time and Space (Exploratory Stage)



Shurui Zhou @2019

Machine Learning: it's time to embrace version control [DataOps]

September 2018 · 9 minute read



Checkpoint 1

use_pretrained = FALSE nepoch=30

train_embeddings = FALSE stop with no improvemnet=5

use_crf = TRUE

use_c

Test

acc

Checkpoint 2

use_pretrained = FALSE nepoch=40

train_embeddings = FALSE stop with no improvemnet=15

use_c

use_c

p=0.5

Checkpoint 3

use_pretrained = FALSE nepoch=40

train_embeddings = FALSE stop with no improvemnet=15

use_crf = FALSE

use_chars = TRUE

Testing model over test set

acc

Checkpoint 4

use_pretrained = TRUE nepoch=40

train_embeddings = FALSE stop with no improvemnet=15

Shurui Zhou @2019

" If you were to map this onto a traditional git workflow, what you would get is thousands of orphaned branches with one or two commits. Which isn't really useful, because none of our UIs are built for tracking thousands of branches, along with the results of those experiments."

https://www.reddit.com/r/MachineLearning/comments/9gakdd/ml_people_are_bad_at_version_control_d/

For AI System, it is hard to do version control

- Code

  + hyperparameter + model + various format of data sets (binary file, data base, etc) + infrastructure

- Large file size (git-lfs , s3)
- Huge configuration space
- Computational expensive
- Cross reference
- Data dependency
- ...

Technical Debt

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Current practices/tools

Meet Michelangelo: Uber's Machine Learning Platform

Jeremy Hermann and Mike Del Balso

September 5, 2017

ModelDB: A system to manage machine learning models

Companies often build hundreds of models a day (e.g., churn, recommendation, credit default). However, there is no practical way to manage all the models that are built over time. This lack of tooling leads to insights being lost, resources wasted on re-generating old results, and difficulty collaborating. ModelDB is an end-to-end system that tracks models as they are built, extracts and stores relevant metadata (e.g., hyperparameters, data sources) for models, and makes this data available for easy querying and visualization.

Use Cases

- Tracking Modeling Experiments
- Versioning Models
- Ensuring Reproducibility
- Visual exploration of models and results
- Collaboration

DVC

Open-source
Version Control System
for Machine Learning Projects

missinglink.ai

DevOps for deep learning

- Run and compare hundreds of experiments
- Version control data in the cloud

Engineering? Education/Tech transfer? Research?

Configuration management for AI systems/pipelines

Version control pipeline for data science