First slide

Good afternoon everyone. Thanks for turning up for my talk in one of the last talks of this Conference.

I am Kurian and today I will be talking about ML models and Dataset versioning. I feel it is an important topic for ML practioners

Second slide

So a bit about myself.

- I am an Open source contributor at DataVersion Control and contributed by writing kaggle kernels about DVC . Also contributed to their docs, Python library and other projects like CloudCV, FOSSASIA too.

- I was a FOSSASIA Open Tech-nights Winner for being an awesome open source contributor.

- I am a Kaggle Expert in Kernels and have finished in Top 5% in some competitions.

- Yeah this is where I am still !

\*Startup Adventures\*

When I was interning in a startup called Neuroplex which was building products to track and make gas stations efficient for IOC. I was working on some interesting computer vision problems mainly on Object detection. Yet I felt a disgusting feeling for Machine Learning there.

As an Open source contributor I was someone who judged amount of work I did per day based on the number of git commits. I was not able to use git initially because of multiple challenges.

\*Challenge 1: ML is slow\* ->

When I was working there, we didn’t write any new code. As most of the code was available online as Open source software. Yet when I wanted to train on a dataset using this code. It took sometimes hours to even 2-3 day to get a decent accuracy for models.

All this time when my model was running, I could have some Tea, snap, fight with my collegues etc

\*Challenge 2: Working with ML projects\*

In software engineering the usual workflow is:

```

git clone <project repo url>

Install requirements

```

Yet when you are working on a datascience github project. You first need to download the dataset and then go on and download pre trained models. If you just want to do interference. Else you want to train your code on a new dataset, it will take lot of time to generate models

data + Code+ models

\*Challenge3: Metric driven\*

Whenever we are working on any machine learning problem, our end goal is usually to create a Model according to a metric. For example a model which classifies flowers in 90% accuracy, generates audio with a particular MOS, for evaluating text output. We use metrics

Whereas in software engineering, we cumulatively build more and more features for our products

/Most used tool for metric tracking is now Excel spreadsheets/

\*Andrew NG Slide\*

ML is not yet efficent. We still don’t have any version control system. Agile might/might no be the solution, then new practises which are yet supported by the tool sets

\*Problems with git and git-LFS\*

After some googling, I realized no one is ever using git in their ML setup and even if they are using it. It’s usually single commit projects which doesn’t track all the aspects of the program.

11 th slide: git excellent tool for code versioning build by Linus Torvalds.

12 th slide:

git can’t handle large files like Datasets and models.

According, it’s recommended to keep github repos to a size of 1 GB.

So to single commit projects, I just want to say you are in this situation:

14 slide:

Git LFS, its another popular tool to handle large files. We use a server to store large data files in service like github, gitlab, atlassian etc.

15th slide:

A few - it requires a server that supports it and it usually means that it's not compatible with regular S3/Google Storage/. In case you run the server yourself (e.g Gitlab) it usually means that you are limited in terms of space (versus using S3, etc). Almost all Git hostings have certain limits - Github - 2GB per file.

also, since it has a very tight integration with Git (you don't run any additional commands) it's not granular enough

with DVC you can pull/push a specific file only

\*How I discovered DVC\*

I was curious and searching tools then I stumbled upon a organisation called DVC(data version control) when I was checking about Google Season of Docs. I very warmly in their community and contributed to their project for a few weeks. Only after a weeks, I realized their tool was solving the problem I was facing while I am interning in Neuroplex.

\*About DVC\*

DVC is a brainchild of a data scientist and an engineer, that was created to fill in the gaps in the ML processes tooling. It’s motto is “For the Data scientist by the datascientist”. DVC adopts the best practices for Machine Learning and make a git like CLI tool for this.

\*Best Practises in ML\*

These are best practises which has been agreed by lot of companies and people in general. And DVC fares relatively good in this criteria.

- So one of problem is versioning ML

- Versioning datasets

- versioning Pipelines

- connecting Data with code

- Tracking metrics

- Visualizing metrics

\*What is Model versioning?\*

Machine Learning is metric driven. Usually the generated models are versioned for each experiments which are performed on this.

\*Tracking experiments\*

It’s very important to keep track of your experiments. Usual what most of researcher when they face a problem, which give not the needed result. They check why is it like that? I can lead to new intuitions and new ideas for the humanity. Like how Resnet blocks originated.

Why versioning?

To keep track of experiments

Monitor both succesful and failed experiments(which happen 75% of time)

\*What is Dataset versioning?\*

Keeping track of versioning is very important. Consider the example of a self driving car like Waymo, which generates 4 TB/day. It’s very important to keep track of your experiments

\*Why dataset managements?\*

> Moving Datasets around

You need to move your dataset from your laptops, to servers, to cloud services where you can use GPUs, TPUs. There needs a tool to solve move datsets easily

> If dataset is evolving dataset versioning is very important. At what instance the change is being made and why it’s very important.

- spreadsheets to track dataset for calculating accuracy

\*\*Use case: Versioning Cats vs Dogs\*\*

- In this problem we are going to Classical deep learning problems of identifying cats vs Dogs

\*Demo Comments:\*

\*Conclusion\*