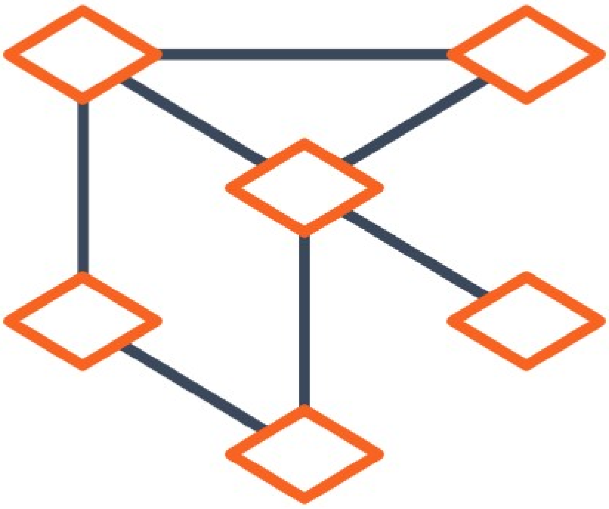
TFX

ML Metadata: Version Control for ML

January 08, 2021

Posted by Ben Mathes and Neoklis Polyzotis, on behalf of the TFX Team

Engineers at Google have learned, through years of hard-won experience, that this history and lineage of ML artifacts is far more complicated than a simple, linear log. You use Git (or similar) to track your code; you need something to track your models, datasets, and more. Git, for example, may simplify your life a lot, but under the hood lineage of your entire ML workflow. Full lineage is all the steps from data ingestion, data preprocessing, validation, training, evaluation, deployment, and so on. MLMD is a standalone library, and also comes integrated in TensorFlow Extended. There's also a demo notebook to see how you can integrate MLMD into your ML infrastructure today.



Beyond versioning your model, ML Metadata captures the full lineage of the training process, including the dataset, hyperparameters, and software dependencies.

Here's how MLMD can help you:

- If you're a ML Engineer: You can use MLMD to trace bad models back to their dataset, or trace from a bad dataset to the models you trained on it, and so on.
- If you're working in ML infrastructure: You can use MLMD to record the current state of your pipeline and enable event-based orchestration. You can also enable optimizations like skipping a step if the inputs and code are the same, memoizing steps in your pipelines. You can integrate MLMD into your training system so it automatically creates logs for querying later. We've found that this auto-logging of the full lineage *as a side effect of training* is the best way to use MLMD. Then you have the full history without extra effort.

MLMD is more than a TFX research project. It's a key foundation to multiple internal MLOps solutions at Google. Furthermore, Google Cloud integrates tools like MLMD into its core MLOps platform:

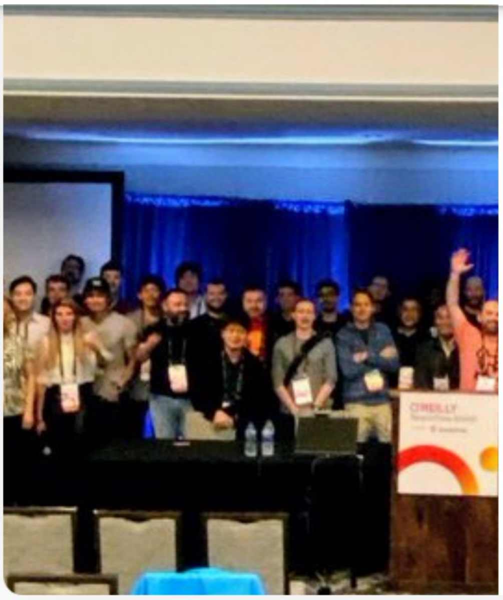
This will enable customers to determine model provenance for any model trained on AI Platform for debugging, audit, or collaboration. AI Platform Pipelines will automatically track artifacts and lineage and AI teams can also use the ML Metadata service directly for custom workloads, artifact and metadata tracking.

Want to know where your models come from? What training data was used? Did anyone else train a model on this dataset already, and was their performance better? Are there any tainted datasets we need to clean up after?

If you want to answer these questions for your users, check out MLMD on github, as a part of TensorFlow Extended, or in our demo notebook.

TFX

Next post



Join the TensorFlow Special Interest Groups (SIGs)

December 22, 2020 — Posted by Joana Carrasqueira, TensorFlow Program Manager and Thea Lamkin, Open Source Program Manager, in collaboration with…