Learnings from Building Mission-critical AI Systems

Abstract

Artificial Intelligence (AI) systems serving traffic in production at scale for complex problems can rarely be accomplished by a single machine learning (ML) model. Instead, one almost always requires a scalable, reliable and resilient service that orchestrates not one, but multiple models working in tandem.

This talk derives from the author and his team's rich experience in building complex mission-critical AI systems for document understanding and comprehension at Intuit. These AI systems classify and extract data from key financial documents at scale. This talk elucidates three important topics to consider while building such AI systems: (1) effective ML project planning & execution, (2) ML models & service evaluation, and (3) seamless upgrades to active in-production systems to manage stakeholder expectations.

Building an AI production system has a key challenge - the probabilistic nature of ML models interacting with the service. The probabilistic nature applies to both the output of a ML model and the model development process. Successful execution requires effective planning and resource estimation under considerable ambiguity and uncertainty. This extends to data scientists building standalone models that solve different sub-parts of a problem, and engineers building the service that deals with compounded complexities that arise from the orchestration of all the models. Creating a plan and an effective design that is extensible, scalable and compartmentalized (using microservices) is paramount before the first line of code is written. It is also foundationally important to evaluate and certify not just individual models but also the service as a whole. Finally, building a well designed and well tested system is less than half the work - the majority of an AI system's life cycle deals with continuous monitoring and maintenance for years, along with having a gameplan to continuously upgrade the same to meet evolving stakeholder needs.

Potential Discussion Points

The goal of this talk is to share key learnings from managing a ML project focused on building a mission critical AI service that orchestrates multiple ML models. This talk will also share experiences, pitfalls and insights from upgrading a seasoned, production-tested AI system serving millions of customers at scale. Some key discussion points of the talk include:

1. Features of a successful AI service development life cycle
    a. Effective planning and effort estimation

    b. Designing a microservices based orchestration service
    c. Effective unit, integration and performance tests
    d. Operations excellence and effective monitoring
2. Evaluating the quality of an AI system
    a. Evaluating models individually
    b. Evaluating the end to end service by evaluating all participating models in tandem
    c. Engaging with stakeholders to relate service performance to stakeholder benefit
3. Rapid reprioritization and execution under time constraints
    a. Dealing with curve balls
    b. Data driven decisions to pivot
    c. Striving for frequent small releases
    d. Ensuring release over release service stability, quality and improvement
4. Key questions to answer when upgrading a tried and tested production system
    a. How to measure improvement against older version
    b. What is the release strategy
    c. What is the rollback plan
    d. If rollback is necessary, can we improve older service mid-season if required

Relevance to Workshop

This talk is highly relevant to the Managing Machine Learning Projects track of the workshop. The document understanding and comprehension case study discussed not only presents useful learnings from managing a ML project, but also shares key insights on multiple aspects of this track such as: case studies and evaluation, collaboration with product development and stakeholders, agile data science, research management, and integration of ML solution in organization.

Presenter Bio

Nandish Jayaram is a Senior Data Science Manager at Intuit, part of the AI, Data and Analytics group (A2D). As part of A2D at Intuit, Nandish leads a team of data scientists and machine learning engineers to build solutions for document understanding and comprehension. The solutions built by his team are used to automatically extract key information from financial documents across TurboTax and QuickBooks suites of products.
Nandish holds a PhD from the University of Texas, at Arlington, where he researched on improving the query systems for large knowledge graphs. Prior to joining Intuit, he was a staff research engineer at Pivotal (now VMWare) and an Apache committer contributing to Apache MADLib - a suite of machine learning algorithms that executed at scale on distributed databases.

Company Portrait

Intuit is the global financial technology platform that powers prosperity for the people and communities we serve with Intuit TurboTax, Credit Karma, QuickBooks, and Mailchimp.Consumers use products like TurboTax and Credit Karma to file taxes, manage finance and more. Small businesses and self employed individuals use products such as QuickBooks and MailChimp to grow their businesses, find and keep customers, pay their employees and more.