netidee

STIPENDIEN

A Serverless Platform for Automated Machine Learning

Zwischenbericht | Call 16 | Stipendium ID 5884

Inhalt

1 Einleitung

Während der letzten Monate gab es sowohl viele Fortschritte als auch einige Änderungen und Knackpunkte innerhalb der Arbeit. Das Ziel der Arbeit bleibt weiterhin wie gehabt, eine Plattform für die Entwicklung von Machine Learning Applikationen zu entwickeln. Hinzugekommen ist, dass nun ein besonderer Fokus auf die automatische Selektierung der Machine Learning Modelle gesetzt wird. Das Platzieren der Modelle selbst wird nun als gegeben angenommen um die Selektierung besser zu analysieren.

2 Status

2.1 **Meilenstein 1 –** State-of-the-Art Recherche

Ziel der Recherche war es, zu verstehen welche existierenden Lösungen es im AutoML und Serverless Bereich schon gibt. Die erreichten Ergebnisse sind im ersten Blogeintrag schon zu sehen. Es wurden sowohl akademische als auch Industrielösungen in diesem Bereich gefunden, welche zunächst sehr hilfreich schienen, aber leider in der langfristigen Betrachtung eher ein Hindernis dargestellt haben.
In der Recherche wurde Amazons SageMaker Plattform stärker beleuchtet, dessen Open Source Container ursprünglich als Ausführungsplattform dienen sollte. Das hat sich weit später als Fehler herausgestellt, da viele Teile der Ausführungscontainer kaum bis gar nicht dokumentiert waren, obwohl diese mit einer Open Source Lizenz ausgestattet waren.
Ein Vorteil, der von der Recherche jedoch hervorgegangen ist, ist dass bestimmte Designziele in späterer Linie getroffen werden konnten, welche nun ein wichtiger Bestandteil der Arbeit sind.

2.2 **Meilenstein 2 –** Plattform-Design und Prototyp-Entwicklung

In diesem Schritt sollte basierend auf dem Wissen der Recherche eine Architektur entwickelt werden, welche in das bisherige Umfeld passt. Aufgrund der damaligen Begeisterung mit SageMaker und der angebotenen Open Source Container zur Ausführung von Machine Learning Modellen wurde die Entscheidung getroffen diese als zentralen Bestandteil der Architektur zu verwenden. Durch die fehlende Dokumentation kam es bei der fortführenden Entwicklung jedoch zu Komplikationen. Insbesondere war es schwer, verschiedene passende Modelle für diese Plattform zu finden und letzten Endes kam die Entscheidung diesen Pfad aufzugeben. Diese Entscheidung kam auch deswegen, da die Erkenntnis kam, dass es nicht zielführend war, monatelang Arbeit zu investieren die Ausführungsplattform lauffertig zu bekommen, statt die gesamte Plattform zu entwickeln. In späteren Schritten wurde dann Tensorflow Serving als Ausführungsplattform gewählt, da die Dokumentation trotz einzelner Mängel hilfreicher war als für den SageMaker Container.

Dementsprechend wurde der Fokus der Prototypisierung auf die ganze Architektur gelegt, statt nur auf die Ausführungsplattform. Dieser Fokus führte zu einem sehr interessanten Open Source Projekt zur Entwicklung von Machine Learning Architekturen: Ray.io. Mithilfe dieser Plattform wurde der erste End-to-End-Prototyp mit automatisierter Modelselektierung entwickelt. Im Laufe dieser Entwicklung wurde auch klar, dass die Modelselektierung allein als Problem schon ausreichend Evaluierungspotenzial für eine Diplomarbeit bietet. Dadurch verschiebt sich der ganze Fokus der Arbeit ein System zur Modelselektierung für Serverless AutoML Plattformen zu entwickeln.
Da nun das Ziel insgesamt exakter war, wurden Nutzungsszenarien entwickelt, die die Erfüllung dieses Ziels testen können. Diese Nutzungsszenarien erfordern eine bestimmte Testumgebung, um verschiedene Hosts mit unterschiedlichen Latenzzeiten zu simulieren. Da die Distributed Systems Group an der TU Wien über einen Kubernetes Cluster zum Testen Verteilter System verfügt, war es ein logischer Entschluss diesen zu verwenden.
Kubernetes war für mich zu diesem Zeitpunkt komplett neu und ich musste mich erst in dieses Thema einarbeiten, was natürlich auch ungeplante Zeit erforderte. Um die Komplexität gering zu halten, wurde Ray.io wieder aus dem System entfernt, da es eine extra Abstraktionsschicht über Clustermanagement Software (wie z.B. Kubernetes) ist. Durch die modulare Gestaltung konnten aber wesentliche Bestandteile beibehalten werden, um die Konsolidierung des Prototyps durchzuführen.

Durch die verschiedenen Änderungen, die im Laufe dieses Meilensteins geschehen sind, wurde bislang kein Blogartikel hierzu veröffentlicht. Dieser wird im September 2023 nachgereicht. Auch der Zwischenbericht hätte in diesem Zeitraum geschehen sollen. Dieser wurde aus den gleichen Gründen versäumt und wird deshalb hiermit nachgereicht.

2.3 **Meilenstein 3 –** Prototyp-Konsolidierung

Für diesen Meilenstein wurden die einzelnen Komponenten der Plattform zusammengeführt in ein Kubernetes-Projekt mit verschiedenen Komponenten. Hier kam schnell die Erkenntnis, dass es sehr wichtig ist, dass die Komponenten leicht und schnell hochgefahren und ausgetauscht werden können, um einen schnellen Entwicklungszyklus zu ermöglichen. Das erforderte zwar zusätzlichen Aufwand, hat sich jedoch gelohnt da die Konsolidierung des Prototyps viel problemfreier verlief als das vorhergehende Architekturdesign und die initiale Entwicklung. Einzelne Probleme konnten sehr schnell gelöst werden und benötigten keine kompletten Neuschreibungen mehr.

Ein Problem gab es bei der Konsolidierungsphase trotzdem. Der bislang eingeplante Kubernetes-Cluster hat sich aufgrund bestimmter Netzwerkeinstellungen in den Nodes nicht für das entwickelte Projekt geeignet, da der Kubernetes-eigene DNS nicht verwendet werden konnte. Dadurch musste auf den Servern der DSG ein neuer Cluster eingerichtet werden, welcher die im Projekt verwendeten Features unterstützt. Als das Projekt letzten Endes lauffähig auf dem neuen Cluster war, starteten die Planungen für die Evaluierung der Plattform.

Der fließende Übergang in die Evaluierung hat mich verhindert den Blogeintrag für diesen Meilenstein zu veröffentlichen und deswegen wird dieser ebenso wie der vorherige im September 2023 nachgereicht. Die Inhalte dazu sind schon vorhanden, da ich die Konsolidierung bereits in der geschriebenen Arbeit dokumentiert habe und Teile davon (mit ein wenig Aufbereitung) leicht verfügbar machen kann.

2.4 Meilenstein 4 - Prototyp-Evaluierung

Mit dem konsolidierten Prototyp im Kubernetes-Cluster wird eine Reihe an Experimenten und Benchmarks durchgeführt, welche Aufschluss über die Leistung der Modelselektierung liefern sollen. Hierbei war es schnell klar, dass eine effektive Evaluierung nur möglich ist, wenn eine breit gefächerte Reihe an Modellen mit verschiedenen Leistungscharakteristika (unterschiedliche Ausführungsdauer und Ergebnisgenauigkeit) vorhanden ist. Die Wahl auf Tensorflow Serving zur Ausführung (statt zuvor der SageMaker Container) war im Grunde sehr gut, da in Tensorflow Hub viele verschiedene Modelle zur Verfügung stehen.
Nachdem passende Modelle gefunden wurden, war aber schnell klar, dass eine effektive Einbindung für alle Modelle sehr lange dauern kann. Deswegen wurde extra für die Evaluierung ein Dummy-Modell entwickelt welches Leistungscharakteristika simulieren kann. Zusätzlich zu diesem Dummy-Modell wurden auch simplere Selektierungsalgorithmen als Vergleichswerte entwickelt.

Beim Starten der Experimente wurde aber schnell festgestellt, dass es noch einige Randfälle gab, in denen das bislang entwickelte Projekt Fehler verursacht hat. Das hat eine bisherige automatisierte Ausführung der Experimente verunmöglicht. Diese Fehler wurden im Laufe der Zeit behoben, aber haben zu einem weiteren Zeitverlust geführt. Dadurch werden erst bis zum Ende vom September 2023 Experimente durchgeführt und evaluiert werden.

2.5 Meilenstein 4 – Finale Verfassung der Arbeit

Durch die vorhergehenden Schwierigkeiten wurde bislang die Finalisierung der schriftlichen Arbeit noch nicht begonnen. Da aber einzelne Teile der Arbeit bereits im Laufe der Entwicklung nebenbei verfasst wurden, wird die Finalisierung weniger dauern als initial geplant. Dementsprechend wird eine Finalisierung der schriftlichen Arbeit in den Monaten Oktober und November 2023 eingeplant.

3 Zusammenfassung Planaktualisierung

Alle Anpassungen des Planungsdokuments kurz zusammengefasst

Aufgrund der genannten Schwierigkeiten in der Evaluierungsphase (und vorhergehenden Fehleinschätzungen) und einer unvorhergesehenen noch andauernden Sehnenscheidenentzündung wird sich die Fertigstellung der Arbeit um 1-2 Monate verzögern. Es wird deshalb um eine Fristverlängerung von 2 Monaten bis Ende November 2023 gebeten. Diese

Verlängerung wird genutzt werden, um den schriftlichen Teil der Arbeit zu vervollständigen. Bis zum Ende September 2023 werden alle Teile der Arbeit, die für das restliche Schreiben der Arbeit notwendig sind, fertiggestellt. Ein erster Vorentwurf wird Mitte Oktober fertiggestellt, welcher bis Ende November schrittweise korrigiert wird.

Die genannten versäumten Blogeinträge werden im September 2023 nachgereicht.