2020-2021

Zuyd Hogeschool
ICT Academie

DISTRIBUTED MACHINE LEARNING

EXTENDED ABSTRACT

Opdrachtgever
Shannen Dolls & Jim Bemelen

Module
BD02

Groepsleden
Bjarne Krottjé
Rianne Ritzen
Martijn Roks

Versie
1.0

Abstract

In dit document wordt het Distributed Machine Learning project beschreven. Dit project is uitgevoerd door Bjarne Krottjé, Rianne Ritzen en Martijn Roks van Zuyd Hogeschool, in opdracht van het lectoraat Data Intelligence. Het doel van dit project is om inzicht te krijgen in de werking van een distributed machine learning systeem. Er is onderzoek gedaan naar wat een distributed machine learning systeem is, welke technieken er worden gebruikt en hoe dit wordt toegepast in de code. Tijdens het onderzoek is er naar voren gekomen dat het doel van het project eigenlijk is om een federated machine learning systeem te ontwikkelen. Dit heeft ertoe geleid dat er een Proof-of-Concept is ontwikkeld dat de werking hiervan aantoont.

1 Inleiding

1.1 Opdrachtgever

Het lectoraat Data Intelligence bestudeert de mogelijkheden en nieuwste inzichten in de wereld van Data Intelligence. Het lectoraat bedenkt oplossingen op het gebied van slim omgaan met data en werkt aan maatschappelijke projecten die Limburg sterker maken. Ook levert het lectoraat praktische én wetenschappelijke inzichten over alles wat met Data Intelligence te maken heeft. (Plan van Aanpak, 2020)

1.2 Aanleiding

De aanleiding van het project vloeit voort uit een ander project van een van de opdrachtgevers. Shannen Dolls, is bezig met het ontwikkelen van een AI-kompas. Dit kompas geeft, aan de hand van een geschetste situatie, een juist advies op het gebied van Artificial Intelligence. (Plan van Aanpak, 2020)

2 Opdrachtomschrijving

2.1 Onderzoeksvragen

Doormiddel van de gesprekken met de opdrachtgever, zijn de volgende hoofd- en deelvragen tot stand gekomen (Onderzoeksrapport, 2020):

- *Hoe werkt een distributed machine learning systeem?*

Doormiddel van de volgende deelvragen probeert de projectgroep de hoofdvraag te beantwoorden:

- Wat is een distributed machine learning systeem?
- Wat zijn voor- en nadelen van een distributed machine learning systeem?
- Welke technieken zijn er om een distributed machine learning systeem op te zetten?
- Hoe wordt een distributed machine learning systeem toegepast?
- Welke dataset wordt gebruikt om het machine learning systeem te trainen?

2.2 Doelstelling

Het doel van dit project is om inzicht te krijgen in de werking van een distributed machine learning systeem. Door hier een onderzoek naar te doen, kan deze techniek ook binnen andere projecten worden toegepast, zoals het AI-kompas project van Shannen Dolls. (Plan van Aanpak, 2020)

2.3 Complexiteit

De complexe taak van dit project is ervoor zorgen dat het machine learning systeem niet gebonden is aan een specifiek apparaat, maar op een federated manier wordt opgezet. Federated machine learning maakt het mogelijk om te trainen met gevoelige data van samenwerkende deelnemers zonder de data zelf te delen. Deelnemers aan het federatieve trainingsproces trainen gezamenlijk een model dat zich met elke trainingscyclus verbeterd. Een deelnemer kan elk apparaat zijn dat Machine Learning kan uitvoeren en kan communiceren met een centrale opslag. (Plan van Aanpak, 2020)

2.4 Artefacten

Tijdens het project zijn een aantal beroepsproducten opgeleverd. De projectgroep heeft een Plan van Aanpak geschreven in het begin van het project. (Plan van Aanpak, 2020) In van het Plan van Aanpak wordt de aanpak van het project beschreven en op welke termijn het project uitgevoerd zal worden. Naast het Plan van Aanpak is er ook een onderzoek gedaan naar hoe een distributed machine learning systeem werkt. (Onderzoeksrapport, 2020) Vervolgens is er een ontwerp opgesteld dat de

achterliggende werking van het systeem in kaart brengt. (Ontwerp, 2020) Aan de hand van het ontwerp wordt er een proof-of-concept opgeleverd dat de werking van een federated machine learning systeem aantoont. (Proof of Concept, 2020) Om het project af te sluiten en te presenteren aan externen, is er ook een posterpresentatie gemaakt die een kort maar duidelijk beeld van het project geeft. (Posterpresentatie, 2020) Daarnaast is er een video gemaakt om het Proof of Concept te demonstreren. (YT Demo Video, 2020)

3 Methode

Voor het onderzoek naar hoe een distributed machine learning systeem werkt, is gekozen om een kwalitatief literatuuronderzoek uit te voeren. Door middel van gesprekken met de opdrachtgever zijn de juiste eigenschappen van het project verwerkt en meegenomen in het onderzoek. Vervolgens is er een deskresearch uitgevoerd om meer informatie te vergaren over een aantal zaken met betrekking tot machine learning. (Plan van Aanpak, 2020)

4 Resultaat

4.1 Onderzoek

Om een goed beeld te krijgen van de opdracht, was het noodzakelijk om eerst een onderzoek uit te voeren naar wat een distributed machine learning systeem is en hoe dit soort systemen werken. (Onderzoeksrapport, 2020)

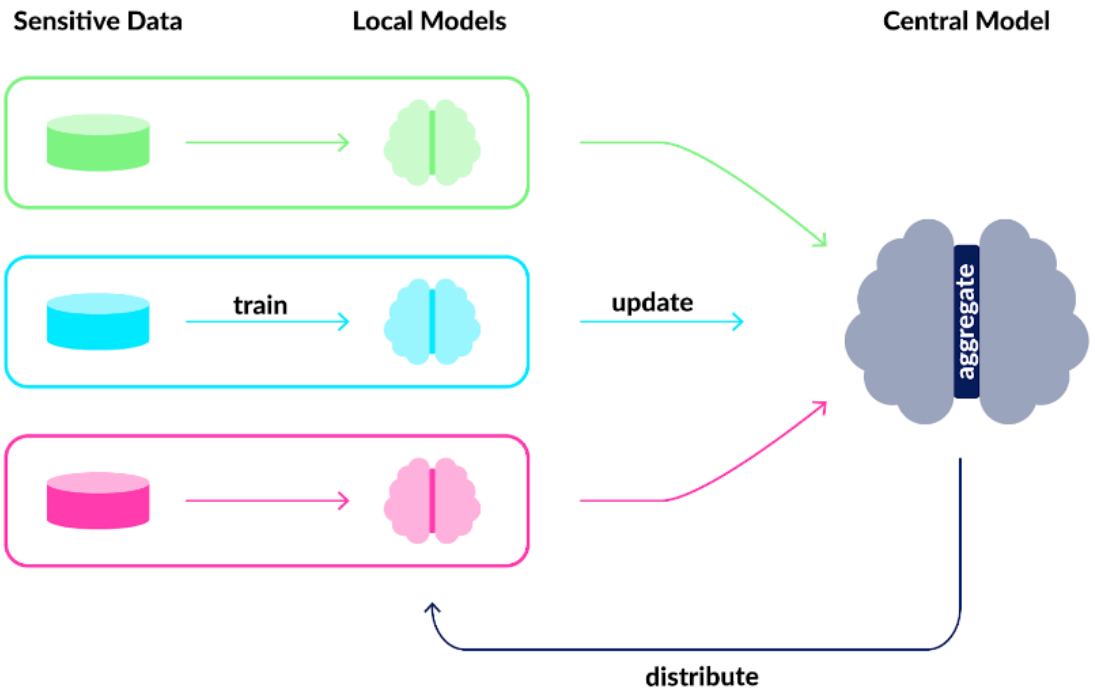
Uit het onderzoek is gebleken dat distributed machine learning gericht is op het distribueren van de trainingslast over verschillende machines of Processing Units. Hierbij staat de data op een centrale plek en verbinden de andere processing units met een centrale processing unit die over de data beschikt.

Om een aantal distributed machines, genaamd werkers, aan elkaar te koppelen, zijn er verschillende technieken beschikbaar. De Ring AllReduce techniek is het meest efficiënte algoritme. Dit komt doordat iedere werker zijn eigen stukjes berekend en alleen het eigen berekende eindresultaat doorstuurt. Hierdoor wordt een eventuele bottleneck voorkomen.

Het machine learning systeem kan opgezet worden op basis van een aantal verschillende platformen. In het onderzoek is naar voren gekomen dat Tensorflow een geschikte optie is om te gebruiken binnen dit project. Tensorflow werkt met Python, een veelgebruikte taal binnen de wereld van Artificial Intelligence. Daarnaast heeft Tensorflow ook een ingebouwde API die ervoor zorgt dat de werkers aan elkaar gekoppeld kunnen worden, genaamd tf.distribute.Strategy.

Echter, is er uit het onderzoek ook gebleken dat de term distributed machine learning niet past binnen de doelstelling van het project. De doelstelling van het project is namelijk om een model te trainen met data die is verdeeld over fysiek gescheiden locaties. De term die hier beter bij past is federated machine learning. Hierbij behouden alle locaties hun eigen data, maar trainen zij gezamenlijk een model dat online wordt opgeslagen.

Wanneer het model is getraind op een van de werkers, wordt het model opgeslagen op een centrale locatie. Hierdoor hebben de fysiek gescheiden locaties allemaal toegang tot het getrainde model. De meest efficiënte manier is dan ook om het model op te slaan in een Azure Blob Storage.



Figuur 1 - Federated Machine Learning

Vervolgens is er samen met de opdrachtgever besloten om bij de kern van het project te blijven en te kiezen voor een lineair regressie model. Dit is om de eenvoud van het model te waarborgen en de focus van het project op het federated uitvoeren te leggen.

Om het federated model te trainen, is er gebruik gemaakt van een supervised training methode. Bij deze methode wordt het systeem eerste geleerd welke data tot welk bijbehorende eindresultaat leidt. Het machine learning systeem berekent het verwachtte resultaat en evalueert hoe ver elke waarde afzit van het werkelijke antwoord.

4.2 Ontwerp

Om een duidelijk beeld van het te ontwikkelen systeem te creëren, is ervoor gekozen om een sequence diagram te maken. In dit diagram is te zien hoe het systeem de verschillende stappen uitvoert van het programma. Hierin staan de stappen van het downloaden van het model tot aan het uploaden naar de Azure Storage. (Ontwerp, 2020)

4.3 Proof-of-Concept

Binnen het project is een eindproduct gerealiseerd dat aansluit op de feedback die de opdrachtgever heeft gegeven tijdens de vergaderingen. Er is een python-applicatie ontwikkeld die de werking van federated machine learning aantoont. (Github Repository, 2020)

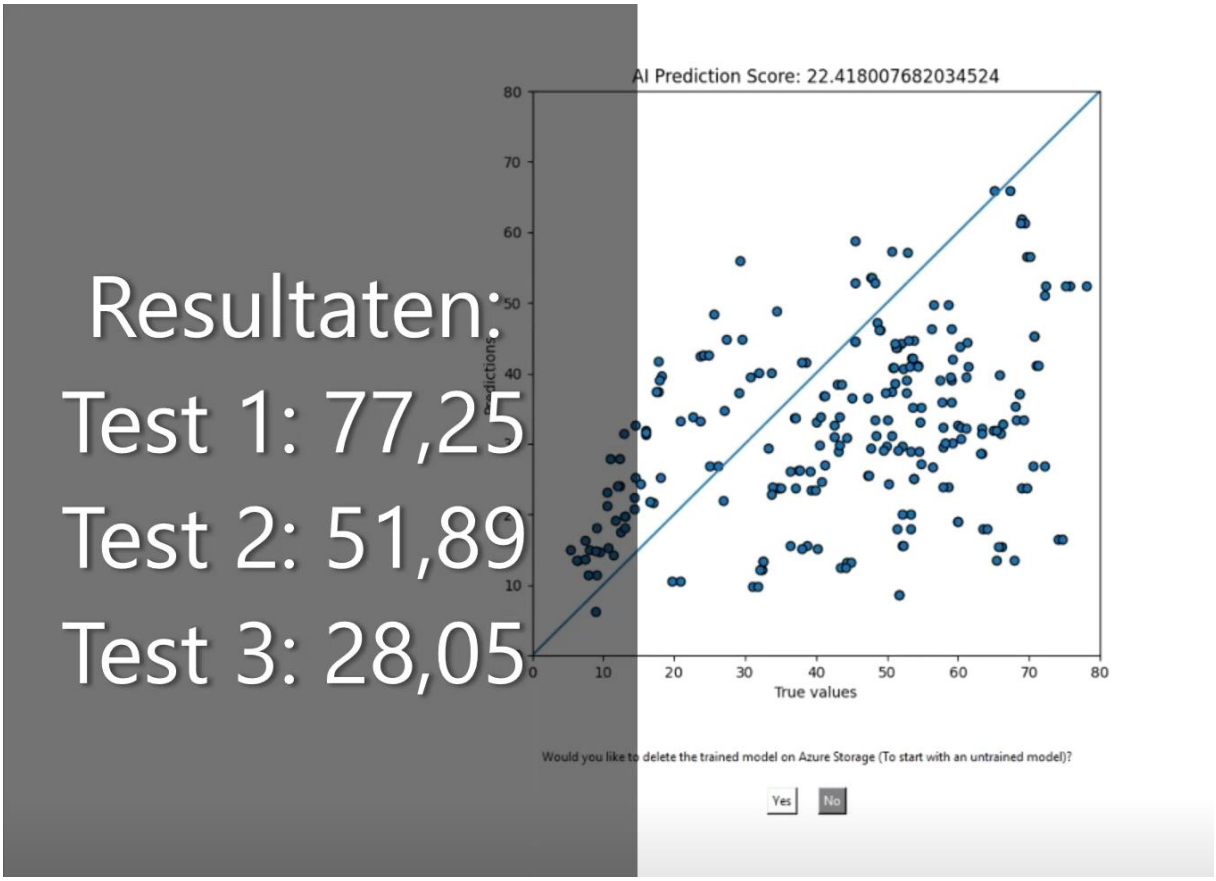
De resultaten van het onderzoek zijn meegenomen tijdens het ontwikkelen van de applicatie. De python-applicatie maakt gebruik van Tensorflow, een veelgebruikt open-source machine learning platform ontwikkeld door Google. Om het model zo simpel mogelijk te houden, is er samen met de opdrachtgever besloten om een lineair regressie model te implementeren.

De applicatie is gekoppeld aan een Azure Blob-storage waar het model wordt opgeslagen. Op het moment dat er nog geen model in de Azure Blob opgeslagen is, maakt het programma een nieuw model aan dat getraind kan worden.

Vervolgens haalt het programma de test- en traindata op uit de map. De namen van deze bestanden kan aangepast worden in de instellingen die gekoppeld zijn aan het programma. Het model wordt getraind en kan vervolgens aan de hand van de test data controleren hoe ver de geschatte resultaten gemiddeld van de werkelijke resultaten afliggen.

Na het testen van de data, wordt het model lokaal opgeslagen. Wanneer er een beschikbare internetverbinding is, wordt het model ook geüpload naar Azure Blob-storage. Hierdoor kunnen andere gebruikers, die eventueel op een fysiek andere locatie aanwezig zijn, het getrainde model ook gebruiken om deze verder te trainen.

Als het model is opgeslagen en eventueel geüpload naar de Blob-storage, geeft het programma de resultaten weer in een grafisch interface (GUI). Hier kan de gebruiker kiezen of ze het model van Azure Blob-storage willen verwijderen of dat ze het model online willen laten staan. Deze functionaliteit is geïmplementeerd om de gebruiker de keuze te geven om opnieuw te beginnen met het trainen. Dit maakt het makkelijker om de functionaliteit van het programma te demonstreren voor een groep aanwezigen. Bij het uitrollen van het machine learning systeem kan deze functionaliteit verborgen worden.



Figuur 2 - Grafisch Interface

5 Discussie

Na een periode van tien weken kan worden geconcludeerd dat het project goed is verlopen en de gewenste resultaten zijn behaald. Gedurende het project zijn er een aantal zaken naar voren gekomen die ervoor hebben gezorgd dat sommige onderdelen van het project meer tijd nodig hadden dan origineel gepland.

De projectgroep had voor de start van het project nog geen kennis of ervaring met betrekking tot machine learning. Dit zorgde ervoor dat er bij het onderzoeken naar mogelijk te gebruiken datasets onvoldoende kennis beschikbaar was om de bruikbaarheid hiervan in te schatten. Hierdoor is er verschillende keren van dataset gewisseld. Dit zorgde ervoor dat de projectgroep pas later aan de slag kon gaan met een correcte dataset.

Daarnaast heeft er op het einde van het project een wijziging in de gebruikte term plaatsgevonden. De opdracht van dit project was namelijk om een distributed machine learning systeem op te zetten. Door de vergaderingen met de opdrachtgever is er naar voren gekomen dat de opdrachtgever een andere bedoeling had voor het project dan opgenomen in de opdrachtomschrijving. Het ging in dit geval niet om distributed machine learning, maar om federated machine learning. Bij distributed machine learning wordt de trainingslast verdeeld over meerdere processing units. Bij federated learning is de data zelf verspreid en trainen de machines los van elkaar, maar trainen wel een gezamenlijk model.

6 Conclusie

Met de huidige resultaten van het project kan de volledige federated machine learning cyclus doorlopen worden. Het getrainde model kan online worden opgeslagen en opgehaald worden door het Proof of Concept. Vervolgens kan het programma het model verder trainen en weer uploaden naar de Azure Blob-storage. Met deze functionaliteiten zijn de zaken die naar voren zijn gekomen in de vergaderingen met de opdrachtgever verwerkt in de eindproducten.

Om een verder vervolg aan het project te geven, is het aan te bevelen om te kijken naar eventuele mogelijkheden om het federated machine learning systeem verder te ontwikkelen. Denk hierbij bijvoorbeeld aan extra functionaliteiten zoals een historie van testresultaten of een verbeterde front-end van de applicatie.

7 Verwijzingen

Ritzen, R., Krottjé, B., & Roks, M. (2020). BRM Groep 11 - BD02 AI - Onderzoeksrapport.

Ritzen, R., Krottjé, B., & Roks, M. (2020). BRM Groep 11 - BD02 AI - Plan van Aanpak.

Ritzen, R., Krottjé, B., & Roks, M. (2020). BRM Groep 11 - BD02 AI - Posterpresentatie.

Ritzen, R., Krottjé, B., & Roks, M. (2020). BRM Groep 11 - BD02 AI - Proof of Concept.

Ritzen, R., Krottjé, B., & Roks, M. (2020). BRM Groep 11 - BD02 AI - YT Demo Video.

Ritzen, R., Krottjé, B., & Roks, M. (2020). Github Repository. *BD02-Distributed-Learning-groep-11*. Opgehaald van https://github.com/ZuydUniversity/BD02-Distributed-Learning-groep-11-

Ritzen, R., Krottjé, B., & Roks, M. (2020). Ontwerp - Readme - Github. Opgehaald van https://github.com/ZuydUniversity/BD02-Distributed-Learning-groep-11-/blob/master/README.md#ontwerp