Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Objektdetektion und Instanzsegmentierung im Edge Computing mit DeepStream SDK und Jetson

Masterarbeit, vorgelegt von Romeo Landry Kamgo Chetchom am 12.04.2021
Studiengang Informatik • Fachbereich Informatik und Medien
Kontak: romeokamgo@gmail.com

Aufgabenstellung

Ein Edge Computing ist eine dezentrale Datenverarbeitung, die aus drei Komponenten besteht. Ein ”IoT Edge Divice” zur lokalen Datenverarbeitung; ein ”Edge Server” oder hoch-virtualisierte Plattform, der zwischen dem Endgerät und dem Server-Cloud Rechenzentrum steht. Ziel der Arbeit besteht darin, eine Architektur für Edge Computing zu implementieren, wobei die Jetson Nano als Edge Gerät benutzt werden kann. Auf der verteilten Architektur soll es möglich sein, moderner Objektdetektionsmodelle einzusetzen. Hierbei sollten mindestens zwei aktuellen Deep Learning Detektoren wie YOLOv3 und SSD verwendet werden. Die Ausführung von Modellen auf dem Edge wurde durch eine geeignete Applikation demonstriert.

Konzept

Die Inbetriebnahme von Modellen auf Jetson Nano wird mit Hilfe von DeepStream SDK gemacht. Die zu ausführende Modelle werden zuerst weiter trainiert, auf der Trainingsplattform zuerst evaluiert und später wird eine weitere Evaluation auf Jetson erfolgen (s. Abb. 1).

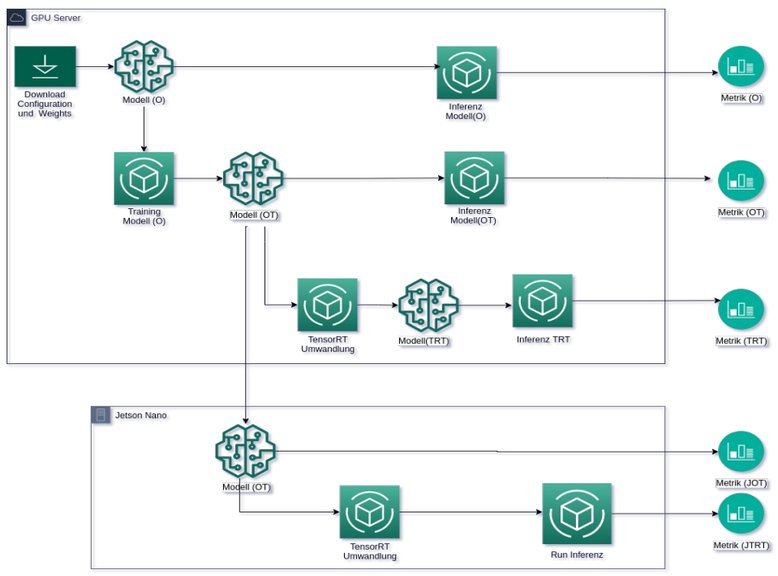

Abb. 1: Vorgehensweise der Analyse der Modelle

Die Modelle, die für die Realisierung dieser Arbeit benutzt wurden, wurden auf Tensorflow2 Model Zoo heruntergeladen. Die Modelle wurde auf COCO Datensatz vor-trainiert und evaluiert. In dieser Arbeit sind die ausgewählten Modelle auf dem gleichen Datensatz (COCO 2017 Datensatz) weiter-trainiert und evaluiert. Neben dem Training und der Berechnung von Metriken finden auch eine Optimierung der Modelle mit TensorRT statt.

Optimierung

TensorRT ist ein Framework zur Inferenz von Modellen. Es beschleunigt den Einsatz des Modells durch eine gewisse mengen von Operationen auf ein trainiertes Modell. Es wurde TensorFlow-TensorRT bei der Optimierung benutzt. TensorFlow-TensorRT ist die integrierte Version von TensorRT in TensorFlow, die dazu hilft, TensorFlow Modelle zu optimieren(s. Abb. 2).

Abb. 2: Ablauf der Optimierung mit TF-TRT

Demo App

Die Anwendung zur Ausführung von Modellen auf Jetson Nano wurde mit Hilfe von DeepStream SDK entwickelt. Dank DeepStream wurde SSD Modell als Basismodell für eine Videoanalyse Applikation entwickelt. Die App benötigt Videostream von einer Kamera als Eingabe. Diese Eingabe wird dank der DeepStream App verarbeiten und die Box auf das zu erkennende Objekt zeichnen. Das verarbeitete Video wird durch RTSP-Protokoll dem Server weitergeleitet und weitere Informationen dazu sind anhand des Apache Kafka Protokolls dem Server weitergeleitet.

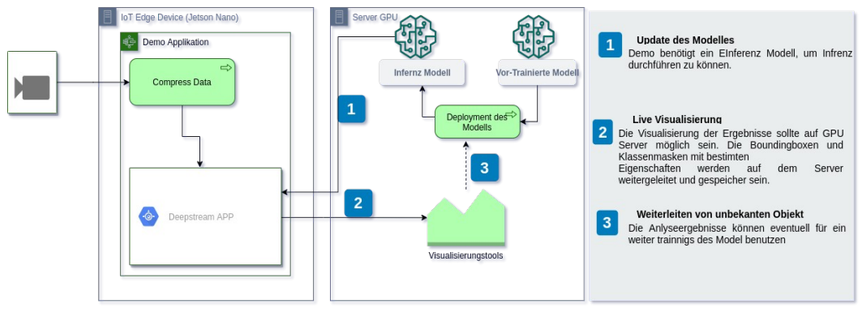

Abb. 3: Deployement von Objekterkennung und Instanz Segmentation Modellen auf Jetson Nano

Ergebnisse

Am Ende dieser Arbeit wurde das SSD-Modell zum Einsatz gebracht. Die Abbildung 4 stellt die Metriken vom Modell SSD dar. Die Mean Average Precision und die Geschwindigkeit des Modells auf dem Server und auf Jetson Nano wurden berechnet.

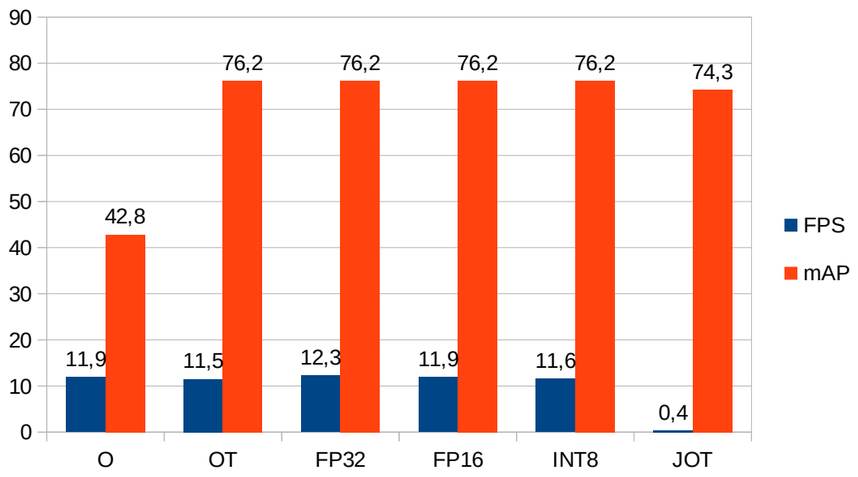

Abb. 4: SSD Metriken

Fazit

In dieser Arbeit wurde einerseits gezeigt, dass die Optimierung von Modellen mit TensorFlow-TensorRT auf Jetson nicht möglich ist, wegen der Inkompatibilität zwischen TensorFlow-TensorRT (Version 2.x) und der vorhandene Version von TensorRT auf Jetson Nano. Anderseits wurde gezeigt, wie man SSD Modell mit Hilfe von DeepStream SDK zum Einsatz bringt.

Betreuung Dipl.-Inform. Ingo Boersch • Technische Hochschule Brandenburg • Prof.Dr.Ing Sven Buchholz • Technische Hochschule Brandenburg