Exposé zur Projektarbeit

Dimensionsreduzierung von KI-Modellen für die Bildverarbeitung (Model-Pruning)

Leandro Carrión Bennewart, Rabea Eschenhagen, Robert Komorowski,
Andrés Eduardo Parilli Fonseca, Sedat Süzer
Betreuung: Tom Wolf

TU Berlin
Institut für Industrielle Automatisierungstechnik
Pascalstraße 8-9, 10587 Berlin

Fraunhofer IPK
Pascalstraße 8 – 9 10587 Berlin

1. Juni 2023

Problemdefintion

Die zunehmende Komplexität von Modellen des maschinellen Lernens stellt eine hohe Anforderung an Ressourcen wie Speicherplatz, Rechenleistung und Energieverbrauch dar. Der Fokus dieses Projektes liegt daher auf der Größenreduzierung von bereits trainierten Modellen durch Model Pruning. Eine Reduzierung der Größe kann dementsprechend zu Vorteilen wie geringerem Speicherbedarf, verbesserten Ausführungszeiten und effizienterer Ressourcennutzung führen. Dies ist insbesondere in Umgebungen mit begrenzten Ressourcen oder auf Geräten mit limitierter Rechenleistung äußerst relevant.

Ziel des Projektes ist es, eine Pruning-Pipline zu entwickeln, die verschiedene Pruning Methoden auf trainierte Modelle anwendet und vergleicht, um das jeweils bestgeeignetste Verfahren zu identifizieren. Dafür wird ein beispielhaftes bereits trainiertes Modell untersucht. Nach der Pruning-Phase, werden die Methoden anhand verschiedener Kriterien, wie Genauigkeit, Ausführungszeit, Speicheranforderungen und Inferenzgeschwindigkeit, bewertet. Die Festlegung von Bewertungskriterien wird dabei einen wichtigen Teil der Evaluierung ausmachen. Eine zusätzliche Zielsetzung besteht darin, den Pruning-Prozess anschaulich darzustellen, um zu einer besseren Verständlichkeit beizutragen.

Unser persönliches Ziel besteht darin, praxisorientierte Erkenntnisse über die Anwendung von Model Pruning zu erlangen und ein breites Verständnis für die verschiedenen Methoden und deren Funktionsweise zu entwickeln.

Vorarbeiten

Um das Konzept Pruning zu verstehen, ist es wichtig, über grundlegendes Fachwissen im Bereich des maschinellen Lernens und neuronaler Netzwerke zu verfügen. Bisher haben wir uns mit folgenden Themen auseinandergesetzt:

1. Überblick über die Themen überwachtes und unüberwachtes Lernen, Optimierungsalgorithmen, sowie Evaluationsmetriken
2. Grundlegende Kenntnisse über Aufbau und Funktionsweise neuronaler Netze.
3. Vertrautheit mit dem Trainingsprozess eines neuronalen Netzwerks, einschließlich Vorwärtspropagation, Rückwärtspropagation (Backpropagation) und Gradientenabstiegsverfahren.
4. Überblick über die gängigen Evaluationsmetriken wie Genauigkeit, Präzision, Recall und F1-Score zur Bewertung der Leistung eines Modells.

Im nächsten Schritt folgt das Aneignen von Kenntnissen über die erforderlichen Schritte für die Durchführung des Pruning-Verfahrens. Dabei geht es darum zu verstehen, wie das Pruning beurteilt und optimiert werden kann, welche Elemente dabei eine Rolle spielen und welche Strategien angewendet werden können, um letztendlich die Leistung und Funktionalität des Modells zu optimieren.

Arbeitspakete

Im Rahmen des Projekts zum Model-Pruning werden folgende Arbeitspakete definiert:

1. Literaturrecherche
    a) Umfassende (individuelle) Recherche zu Model-Pruning.

b) Erstellung einer gemeinsamen Lese-Liste mit den gefundenen Artikeln.

c) Identifizierung und Priorisierung der Hauptquellen.

d) Zusammenfassung der wichtigsten Erkenntnisse und Methoden.

e) Pflegen einer Quellendatenbank.

(Verantwortliche Person: Robert; Durchführende Personen: Alle)

2. Datenvorbereitung
   a) Beschaffung des vortrainierten neuronalen Netzwerks und der erforderlichen Datensätze.
   b) Vorbereitung für das Pruning.

   (Verantwortliche Person: Leandro; Durchführende Personen: Andrés und Leandro)

3. Entwicklung der Pruning-Pipeline
   a) Implementierung verschiedener Pruning-Methoden.
   b) Integration der Methoden in eine Pruning-Pipeline.

   (Verantwortliche Person: Andrés; Durchführende Personen: Andrés und Leandro)

4. Evaluierung der geprunten Modelle
   a) Festlegung von messbaren Leistungsmetriken, wie Genauigkeit und Inferenzgeschwindigkeit.
   b) Erstellung eines Experimenteplans.
   c) Auswertung der Experimente zur Bewertung der geprunten Modelle.

   (Verantwortliche Person: Rabea; Durchführende Personen: Sedat und Rabea)

5. Visualisierung der Ergebnisse
   a) Entwicklung einer Methode zur Visualisierung der Ergebnisse und Vergleich der verschiedenen geprunten Modelle.
   b) Erstellung von Diagrammen, Grafiken oder anderen visuellen Darstellungen, um die Leistungsunterschiede zu verdeutlichen.

   (Verantwortliche Person: Sedat; Durchführende Personen: Robert und Sedat)

6. Erstellung des Docker-Containers
   a) Konfiguration und Erstellung eines Docker-Containers mit der Pruning-Pipeline und den erforderlichen Abhängigkeiten.
   b) Sicherstellung der Portabilität und Kompatibilität auf verschiedenen Systemen.

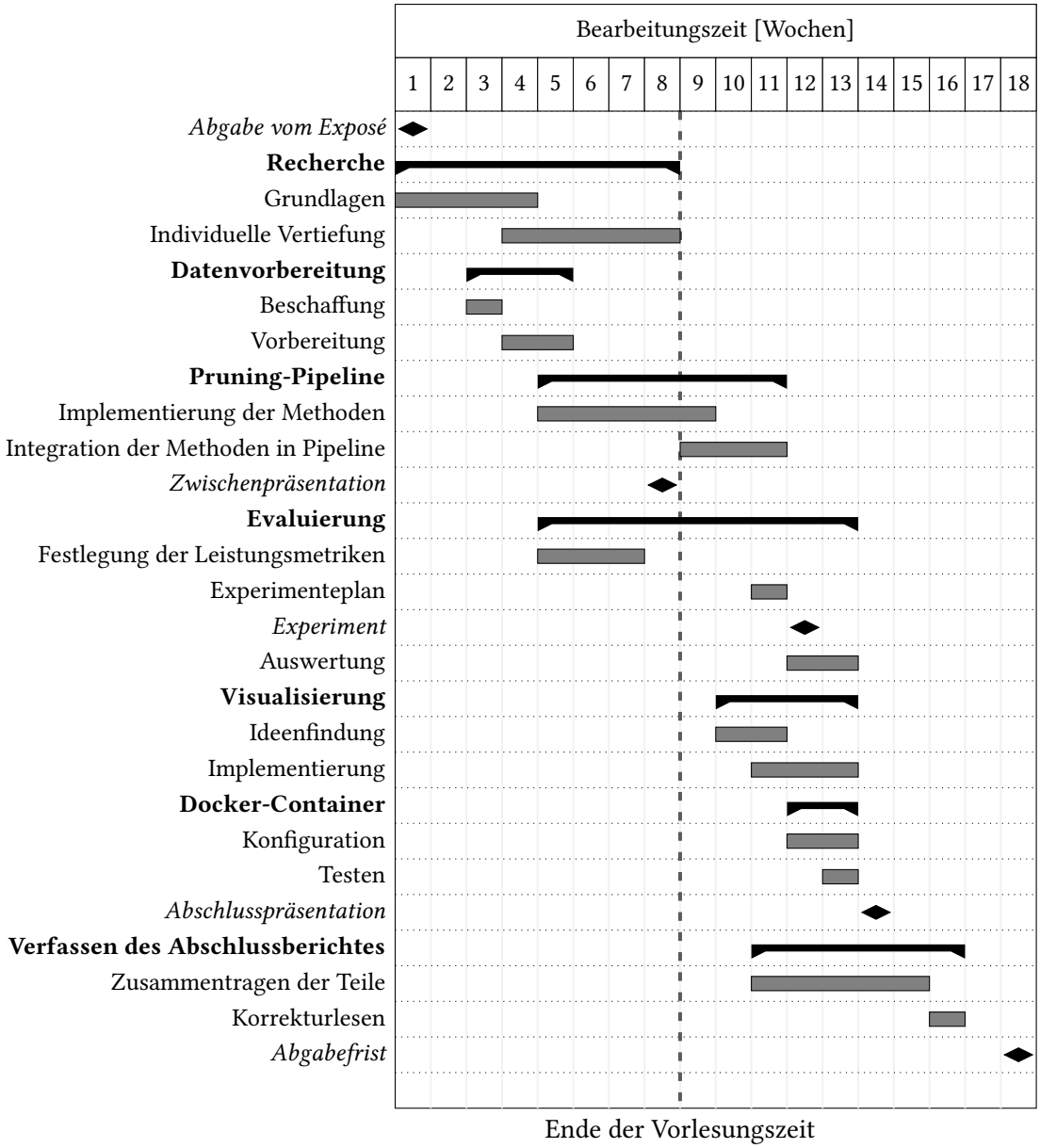
   (Verantwortliche Person: Robert; Durchführende Personen: Robert und Rabea)

7. Verfassung der Ausarbeitung
   (Verantwortliche Person: Rabea; Durchführende Personen: Alle)

8. Ausblick: gegebenenfalls Experimentplanung für die Klassifizierung von Autoteilen mit dem geprunten Modell
   a) Erstellung eines Exeprimenteplans.
   b) Testen der Pipeline am realen System.

Zeitplan

Der folgende Zeitplan erstreckt sich über einen Zeitraum von 18 Wochen, beginnend am 29.5. Um ausreichend Pufferzeit bis zur Abgabefrist am 30.9. zu haben, streben wir an, alle technischen Komponenten des Projekts bis zum 1.9. abzuschließen.



Betreuung und Projektorganisation

Die während des Projektes erzielten Ergebnisse werden dem Betreuer in regelmäßigen Abständen von zwei Wochen in einem Onlinemeeting vorgestellt. Zusätzlich finden wöchentliche Gruppentreffen für die Projektkoordination statt. Alle Projektbestandteile werden in einem gemeinsamen GitHub-Repository gesammelt. Um eine detaillierte Aufgabenverwaltung zu ermöglichen, werden die Arbeitspakete in einem Kanban-Board innerhalb des Repositories festgehalten. Dadurch behalten wir den Überblick über ausstehende Aufgaben und es wird transparent, welches Gruppenmitglied gerade an welcher Aufgabe arbeitet. Für das Verfassen der Texte nutzen wir Overleaf (LaTeX) und führen dort auch eine Quellendatenbank.