Vergleich des Umgangs mit Trainingsdaten und Modellen im Digitalisierungsprozess von historischen Schriften

Leistungsnachweis 1

von

Andrés Baumeler

Telefon: 076 443 04 71, E-Mail: andres@baumeler.dev

Alte Riedikerstrasse 5c, 8610 Uster

Betreuer

Clemens Neudecker

Universität Zürich, Philosophische Fakultät /
Zentralbibliothek Zürich

CAS Datenmanagement und Informationstechnologien

Inhaltsverzeichnis

Kapitel 1

Einleitung

Bei der Digitalisierung von historischen Schriften nimmt Optical Character Recognition (OCR) eine wichtige Rolle ein. OCR wird im Digitalisierungsprozess dazu verwendet Text von digitalisierten Dokumenten maschinenlesbar zu machen. In den letzten Jahren konnte die Qualität der OCR Resultate durch den Einsatz von Maschinenlernen, insbesondere neuronale Netzwerke, verbessert werden. So ist es heute möglich auch handgeschriebenen Text automatisch zu erkennen. Dies ermöglicht die Retrodigitalisierung von historischen Schriften in einem vorher nie dagewesenen Volumen. Der Einsatz von Neuronalennetzwerken setzt voraus, dass entsprechende Trainingsdaten in genügender Qualität und Quantität verfügbar sind um die eingesetzten neuronalen Netzwerke zu trainieren. Das Ergebnis des Trainingsvorgangs wird in Modellen gespeichert. Diese Modelle können für verschiedene Szenarien wiederverwendet werden und ersparen ein erneutes Trainieren des neurolaen Netzwerks. In diesem Text soll anhand zwei Beispielen aufgezeigt werden wie heute mit Trainigsdaten und Modellen im Bereich der Digitalisierung von historischen Schriften umgegangen wird. Dabei werden die zwei Frameworks OCR-D und Transkribus analysiert und verglichen. OCR-D setzt auf einen open Source und verteilten Ansatz während Transkribus auf einen closed Source und zentralisierten Ansatz setzt.

Im Ersten Kapitel wird aufgezeigt wie neuronale Netzwerke für die Texterkennung eingesetzt werden und welche Rolle dabei die verwendeten Trai-

ningsdaten spielen. Im Zweiten Kapitel werden die Frameworks OCR-D und Transkribus vorgestellt. Weiter wird aufgezeigt wie die beiden Frameworks mit Trainingsdaten und trainierten Modellen umgehen.

Kapitel 2

Grundlagen

2.1 Einsatz von neuronalen Netzwerken für OCR

Neuronale Netzwerke (NN) sind ein Subset des Bereichs Maschinenlernen (ML). Neuronale Netzwerke bestehen aus einer Eingangs- und Ausgangsschicht. Dazwischen gibt es eine oder mehrere sog. versteckte Schichten. In jeder Schicht gibt es Knoten (künstliche Neuronen) welche mit einem oder mehreren Knoten aus anderen Schichten verbunden sind. [1].

2.2 Trainieren von neuronalen Netzwerken

Begriffe Trainingsdaten, Ground Truth und Modelle erklären. Aufzeigen wie neuronale Netzwerke trainiert werden und wie die verwendeten Trainingsdaten das Endergebnis beeinflussen können. Aufzeigen wie Trainingsdaten und Modell zusammenhängen. Aufzeigen wie ein Modell aus Trainingsdaten entsteht.

2.3 Handhabung Trainingsdaten und Modelle

Aufzeigen wie Trainingsdaten und Modelle Strukturiert sind. Aufzeigen welche Methoden es heute gibt um Trainingsdaten und Modelle zu verwalten.

Kapitel 3

Vergleich des Umgangs mit Trainingsdaten und Modellen

3.1 OCR-D

Was ist OCR-D, wo kommt OCR-D zur Anwendung?

3.2 Transkribus

Was ist Transkribus, wo kommt Transkribus zur Anwendung?

3.3 Umgang mit Trainingsdaten

Aufzeigen wie die beiden Framworks mit Trainingsdaten umgehen. Wie kann neue Ground hinzugefügt werden.

3.4 Umgang mit Modellen

Aufzeigen wie die beiden Frameworks mit Modellen umgehen. Wie können Modelle wieder verwendet werden. Wie können trainierte Modelle geteilt werden. Welche Daten von einem Modell sind einsehbar. Kann ein Modell weiter trainiert werden.

Kapitel 4

Schluss

Zusammenfassung

Abkürzungen

OCR	Optical Character Recognition
ML	Maschinenlernen
NN	Neuronales Netzwerk

Literaturverzeichnis

[1] IBM. What is a neural network? Website, 2023. URL https://www.ibm.com/topics/neural-networks.