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

Digitalisierung von historischen Schriften ist ein umfassender Prozess welcher aus vielen Schritten besteht. Optical Character Recognition (OCR) nimmt dabei eine wichtige Rolle ein. OCR wird im Digitalisierungsprozess dazu verwendet Text von digitalisierten Dokumenten maschinenlesbar zu machen. Dadurch können die Dokumente im Volltext durchsucht werden und der Text für weitere Analysen verwendet werden. In den letzten Jahren konnte die Qualität der OCR Resultate durch den Einsatz von Verfahren aus dem 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 neuronalen netzwerken 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 neuronalen Netzwerks. Für die Volltexterkennung von historische Schriften haben sich dabei zwei Lösungen heruaskristallisiert: OCR-D und Transkribus. Diese zwei Lösungen, insbesondere der Umgang mit Trainngsdaten und Modellen sollen in diesem Text verglichen werden. 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 die Volltexttransformation abläuft und wo im Prozess Neuronales Netzwerk (NN) eingesetzt werden können. Weiter wird erläutert wie Neuronale Netzwerke trainiert werden und welche Rolle dabei die verwendeten Trainingsdaten spielen. Im Zweiten Kapitel werden die Frameworks OCR-D und Transkribus vorgestellt. Dabei wird aufgezeigt wie die beiden Frameworks mit Trainingsdaten und trainierten Modellen umgehen. In desem Abschnitt wird auch auf die Vor- und Nachteile der beiden Lösungsansätze eingegangen. Es handelt sich hierbei nicht um einen wissenschaftlichen Vergleich sondern um einen Erfahrungsbericht aus der Anwendung der zwei Lösungen im Privatbereich.

Kapitel 2

Grundlagen

2.1 Volltexttransformation

Der Prozess der Volltexttransformation besteht aus mehreren Schritten wovon die eigentliche Texterkennung nur einer ist. Je nach Ausgangslage- und Material sind dabei unterschiedliche Schritte notwendig. Ein Beispielhafter Prozess zur Volltextransformation beinhaltet die Schritte:

- Seitentrennung
- Aufbereiten der einzelnen Seitentrennung
- Optische Layout Erkennung und Segmentierung der Seiten
- Aufbereiten der einzelnen Segmente
- Unterteilen der einzelnen Segmente in Textzeilen
- Aufbereiten der einzelen Textzeilen
- Texterkennung auf den einzelnen Textzeilen
- Zusammenfügen der erkannten Texte
- Klassifizierung der erkannten Regionen
- Dokumentanalyse und Export
- Nachbearbeitung und ggf. manuelle Kontrolle und Korrektur der Resultate

Die Unterteilung in einzelne Textzeilen ist notwendig, da die eingestzten NN die Texterkennung auf Zeilenebene durchführen. Für die meisten dieser Schritte können bzw. werden NN eingesetzt. Der Fokus des Vergleichs wird aber auf den Schritten Segmentierung und Texterkennung liegen.

2.2 Einsatz von neuronalen Netzwerken für Volltexttransformation

NN sind ein Subset des Bereichs Maschinenlernen (ML). Neuronal Netze werden verwendet, um Muster in Daten zu erkennen und dadruch Probleme zu lösen. In den letzten Jahren hat die Entwicklung von neuronalen Netzen enorme Fortschritte gemacht, insbesondere durch den Einsatz von Deep Learning, einer Variante des neuronalen Netzes, die mehrere Schichten von Netzwerken verwendet, um komplexe Aufgaben auszuführen. 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. Jede Schicht verarbeitet Eingabedaten und gibt Ausgabedaten an die nächste Schicht weiter bis die Daten am Schluss auf der Ausgabeschicht landen. [8]

Bei der Volltexterkennung können NN eingesetzt werden um die Layouterkennnung, Segmentierung und Texterkennung der einzelnen Segmente durchzuführen. Dafür sind für jeden Schritt andere Netzwerke mit anderne Trainingsdaten notwendig.

2.3 Trainieren von neuronalen Netzwerken

Trainingsdaten sind eine Sammlung von Beispielen, die zum Training von künstlichen Intelligenz Modellen verwendet werden. Sie bestehen aus Eingabe- und Ausgabedaten. Die Qualität der Trainingsdaten hat einen grossen Einfluss auf die Leistung des Modells, da sie dem Modell beibringen, Muster zu erkennen und Entscheidungen auf der Grundlage dieser Muster zu treffen.

Ground Truth (zu Deutsch: Wahrheit oder Wirklichkeit) ist ein Begriff der in der künstlichen Intelligenz verwendet wird um die korrekten Ausgaben für eine Eingabe zu beschreiben. Ground Truth sind Trainingsdaten für welche die Eingangs - und Ausgangsdaten verifiziert wurden. Das bedeuetet zu einem Eingangswert ist der korrekte Ausgangswert fesgehalten. Bei der Texterkennung können Trainingsdaten beispielsweise aus Bilder von Textzeilen und dem auf dem Bild enthaltenen Text in maschinenlesbarer Form bestehen. Diese Ground Truth muss meist durch mühsame manuelle Arbeit erstellt werden denn nur so kann sichergestellt werden, dass das Modell auf einer korrekten Grundlage trainiert wird.

Im Kontext der neuronalen Netzwerke bezeichnet ein Modell die Konfiguration eines NN. Zur Konfiguration gehören die Anzahl und Anordnung der Neuronen, die Verbindungen unter den Neuronen sowie die Gewichtung dieser Verbindungen. Beim Training werden diese Werte stetig verändert um die Ausgabendaten des NN

Ein Netzwerk wird trainiert, indem es mit Trainingsdaten gefüttert wird um daraus Ausgangsdaten gemäss der aktuellen Konfiguration zu produzieren. Die Ausgabedaten werden mit den Eingabedaten verglichen und das Netzwerk wird angepasst. Im Zuge dieser Anpassung können etwa Verbindungen zwischen den Neuronen gelöscht oder neu angelegt werden. Auch die Gewichtung der einzelnen Verbindungen kann angepasst werden. Diese Anpassungne erfolgen automatisiert und semi-zufällig. Ein Training geht über mehrere Durchläufe (sogenannte Epochen). Es gibt verschiedene Metriken um zu messen wie gut ein Modell an den korrekte Ausgangsdaten ist.

Wenn die Trainingsdaten nicht repräsentativ für das zu lösende Problem sind oder schlecht gelabelt wurden, kann das Modell fehlerhaft trainiert werden und schlechte Vorhersagen treffen. Daher ist es wichtig, qualitativ hochwertige Trainingsdaten zu haben, um ein leistungsfähiges Modell zu trainieren.

2.4 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

OCR-D wird im Rahmen des DFG-Projekts OCR-D entwickelt und hat zum Ziel die Volltexttransformation Drucken aus dem Deutschen Sprachraum des 16. bis 18. Jahrhunderts konzeptionell und technisch vorzubereiten [7]. Das OCR-D Framework ist OpenSource. Das beudeutet der Source Code ist öffentlich einsehbar und kann von interssierten Personen auch modifiziert werden. OCR-D verwendet zur Verwaltung des Source Code ein Repository auf GitHub.[1]

OCR-D ist ein Framwork welches mehrere Softwaremodule verbindet. Durch diese Modulare herangehensweise können die einzelnen Schritte im OCR Workflow druch unabhängige Softwarekomponenten abgedeckt werden. Diese Softwarekomponenten werden im OCR-D Framework als Prozessoren bezeichnet. Für den OCR Prozessor können die OCR Engines Tesseract, Ocropus, Kraken und Calamari eingesetzt werden. Der Export der OCR Resultate geschieht im ALTO Format.

Die Verwendung von OCR-D ist kostenfrei. Unterstützung beim Setup und der Anwendung kann aus der OCR-D Community bezogen werden.

Der Anwendungsbereich für OCR-D sind Institutionen welche in der Lage

sind die Lösung selbst zu Installieren und zu Konfigurieren. Die Installation kann dabei nativ oder via Docker Container erfolgen. Bei der nativen Installation wird das Programm vom Source Code gebaut und installiert. Dies geschieht mit den Werkzeugen make. Bei der Installation via Docker wird eine eine Maschine oder Server mit einer Docker Engine vorausgesetzt. Das Docker image wird dann via docker pull vom Repository geladen. Der Export der OCR Resultat ist als PDF oder im ALTO Format möglich.

3.2 Transkribus

Transkribus ist eine kommerzielle Plattform für Texterkennung, Transkription und das Durchsuchen von historischen Dokumenten. Transkribus wurde im Rahmen des Horizon 2020 EU-Projekts READ von einem Konsortium führender Forschungsgruppen aus ganz Europa unter der Leitung der Universität Innsbruck entwickelt. Die Plattform wird von der Genossenschaft READ-COOP betrieben und weiter entwickelt. [2]

Für die einzelnen Schritte Layout- und Strukturerkennung sowie Texterkennung setzt Transkribus auf Maschinelles Lernen. [3]

Für die Texterkennung muss bei Transkribus bezahlt werden. Die Bezahlung erfolgt mit Credits welche vorgängig über den Transkribus Shop gekauft werden müssen. [6]

Um Transkribus anzuwenden ist keine Installation notwendig. Dokumente können direkt über die Webseite von Transkribus unter https://transkribus.ai möglich. Dabei werden die Bildaten an die Server von Trankribus übermittelt und durchlaufen dort die Schritte für die Volltexttransformation. Für kleinere Mengen an Seiten entfällt somit eine lokale Installation, es wird lediglich ein aktueller Browser vorausgesetzt.

Für grössere Volumen bietet Transkribus auch eine Client Anwendung an. Dieser sogennnante Expert Client ist eine Eclipse basierte Java Applikation welche lokal installiert wird. Über diesen Client lassen sich eine vielzahl Parameter für den Prozess der Volltexttransformation einstellen. Der Client dient auch dazu Bilder zu Transkribus Servern hochzuladen. [5]

Für die Anbindung an andere System beitet Transkribus auch ein REST

API zum Hochladen von Dokumenten und zum konfigurieren der Volltextransformation. [4].

Die Bearbeitung der Bilder erfolgt aber auch in diesem Verfahren auf den Servern von Transkribus.

3.3 Umgang mit Trainingsdaten

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

Bei OCR-D liegt es in der Verantwortung der Anwender die Trainingsdaten zu verwalten.

Transkribus bieted die Möglichkeit eigene Bilder zum Training hochzuladen. Für diese Trainingsdaten können dann Zugriffsrechte vergeben werden. So ist es möglich ein Modell zwar öffentlich zugänglich zu machen, die zugrundeliegenden Trainingsdaten aber privat zu halten. Für die Modelle welche von Transkribus selbst veröffentlicht wurden, sind die Trainingsdaten nicht zugänglich.

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.

Bei OCR-D sind die Modelle welcher erstellt werden immer unter der Kontrolle der erstellenden Institution. Da es sich um Open Source Softeare handelt ist offen einsehbar wie ein Modell zustande kommt. Da OCR-D ein Framwork ist welches verschieden Prozessoren für die Texterkennung einsetzt müssen Modelle für die einzelnen Modelle separat traniniert werden. Die einzelnen Modelle sind untereinander auch nicht zwingendermassen kompatibel. Beispielsweise ist ein Modell welches für die Ocropus OCR Engine trainiert wurde nicht zwingendermassen kompatibel für die Tesseract OCR Engine.

Von diesen Einschränkungen abgesehen, können Modelle aber offen miteinander geteilt werden. Modelle werden als ZIP Container ausgetauscht und können auf einer passenden Umgebung weiter verwendet werden.

Die folgenden Informationen beziehen sich auf den Transkribus Expert Client in der Version 1.24.2. Tranksribus unterscheided bei Modellen zwischen Modellen für die Layouterkennungen und Modellen für die Texterkennung. Bei den Modellen für die Texterkennung können Modelle entweder für die CITLab Engine oder für die PyLaia Engine trainiert werden. Da bei Transkribus die Verarbeitung immer auf der Platform und damit auf den Servern von Transkribus stattfinden, liegen auf trainierte Modelle auf der Platform. Transkribus bietet die Möglichkeit die Zugriffsrechte auf einem Modell anzupassen. So ist es möglich ein Modell öffentlich zu teilen und allen Benutzer der Platform den Zugriff auf das Modell zu ermöglichen. Ein Export des Modells, etwa zur Verwendung in einer eigenen Installation oder als Backup, ist aber nicht möglich. Die Rohdaten zu einem Modell sind nicht einsehbar. Transkribus zeigt auf der Seite zu einem Modell eine Übersicht mit Statistiken und Informationen zu einem Modell. So ist ersichtlich welche Trainingsdaten verwendet wurden und wie viele Epochen trainiert wurden.

Kapitel 4

Schluss

OCR-D und Transkribus lösen ähnliche Probleme mit unterschiedlichen Ansätzen. OCR-D setzt auf Open Source während Transkribus eine mehrheitlich geschlossene Plattform ist.

Die Einstiegshürden für Transkribus sind gerade für Institutionen mit geringem Informatikverständnis um einiges tiefer als bei OCR-D. Bei Transkribus reicht eine Registierung aus um erste Dokumente über den Browserclient zu digitalisieren. Ein initial Setup von OCR-D benötigt mehr technisches Verständnis, bietet dann aber mehr Konfigurations- und Anpassungsmöglichkeiten. Der Vorteil von OCR-D ist, dass sämtliche verwendeten Komponenten Open Source sind. Dadurch kann wird die Abhängigkeit von einer Institution verringert. Der Prozess kann zudem transparent dokumentiert werden indem die verwendeten Prozessoren inklusiver Versionen und Parameter notiert werden.

Bei der Handhabung von Trainingsdaten zeigen sich Unterschiede zwischen den zwei Framworks. Bei OCR-D sind die trainierten Modelle offen zugänglich und können geteilt werden. Hier stellen sich Probleme der Kompatbilität. Modelle eines Prozessors sind nicht unbedingt mit einem anderen Prozessor nutzbar. Auch bei unterschiedlichen Prozessor- oder Engine Versionen kann es zu Inkompatibilitäten kommen. Bei Transkribus existiert zwar die Möglichkeit Modelle zu teilen. Der Zugriff auf die Rohdaten eines Modells ist aber nicht möglich. Das verhindert den Einsatz eines Modells auf

einener anderen Plattform. Zugleich wird durch den Ansatz von Transkribus die Komplexität reduziert. Die Benutzenden brauchen sich keine Gedanken um die Kompatbilität der Modelle zu machen. Transkribus bietet die Möglichkeit Zugriffsrechte für Trainingsdaten und Modelle separat zu kontrollieren. Das ermöglicht es ein Modell zu veröffentlichen auch wenn die zugrunde liegenden Trainingsdaten etwa aus rechtlichen Gründen nicht veröffentlicht werden dürfen.

Beide Plattformen bieten einen guten Funktionsumfang und lieferen beeindruckende Resultate in der Volltexttransformation. Sowohl der Open Source Ansatz von OCR-D als auch der Ansatz einer kommerziellen Plattform bei Transkribus machen Sinn. Beide Ansätze kommen mit ihren jeweiligen Vor- und Nachteilen. Es kann nich abschliessend gesagt werden, welcher der zwei Ansätze besser ist. In der Evaluationsphase eines Digitalisierungsprojektes sollte deshalb sorgfälltig anhand der Anfoderungen des Projekts und der durchführenden Institution entschieden werden welche Lösung die geeignetere ist.

Abkürzungen

OCR	Optical Character Recognition
ML	Maschinenlernen
NN	Neuronales Netzwerk

Literaturverzeichnis

[1] Ocr-d. Git Repository, 02 2023. URL https://github.com/OCR-D.

[2] Wir sind read-coop. Website, 2023. URL https://readcoop.eu/de/ueber-uns/.

[3] Transkribus. Website, 02 2023. URL https://readcoop.eu/de/transkribus/.

[4] Rest-api. Website, 02 2023. URL https://readcoop.eu/de/transkribus/docu/rest-api/.

[5] Transkribus herunterladen. Website, 02 2023. URL https://readcoop.eu/de/transkribus/download/.

[6] Credits und preise. Website, 02 2023. URL https://readcoop.eu/de/transkribus/credits/.

[7] Elisabeth Engl. Ocr-d kompakt: Ergebnisse und stand der forschung in der förderinitiative. *Bibliothek Forschung und Praxis*, 44(2):218–230, 2020. doi: doi:10.1515/bfp-2020-0024. URL https://doi.org/10.1515/bfp-2020-0024.

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