Layout-Erkennung in digtalisierten Dokumenten mittels Neuronaler Netzwerke

Exposé

Jakob Schmolling

28. Dezember 2017

Digitalisierte Dokumentensammlungen

Im Internet hat man heute Zugang zu Millionen digitalisierter Dokumente über Anbieter wie Google Books, archive.org und den Sammlungen unterschiedlister Bibliotheken. Aus einer eingescannten Buchseite können keine Informationen über den Inhalt ausgelesen werden. Erst optische Zeichenerkennung (OCR) macht es möglich Zeichenketten aus den Bilddaten zu gewinnen. Bücher und Magazine sind aber mehr als einfache Zeichenketten und können unterschiedliche “Objekte” wie Illustrationen oder Fotos enthalten. Zum anderen wird Typgraphie meistens eingesetzt um einem Text eine Hierarchie zu geben.

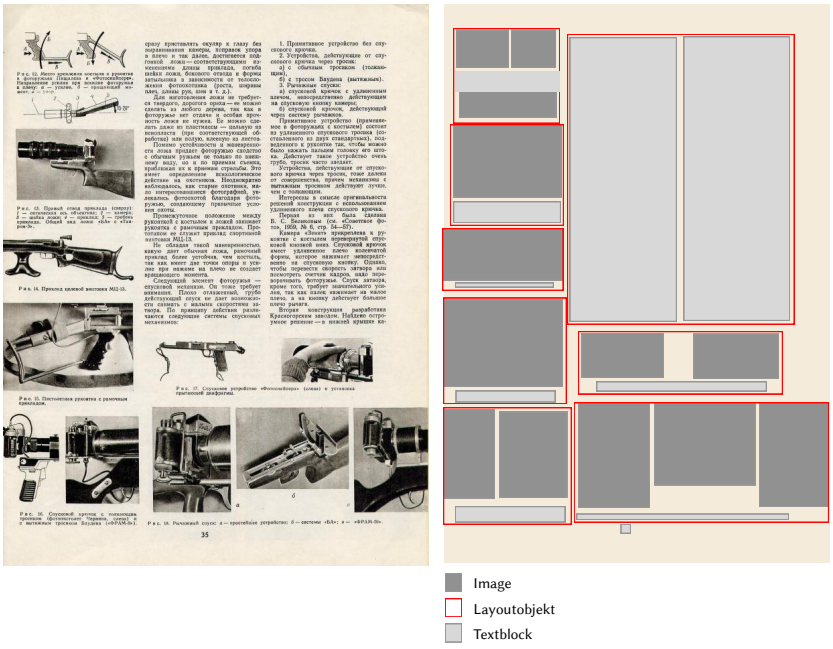

Abbildung 1: (links) Beispielseite aus einem Fotomagazin [__**1967**] (rechts) Eine mögliche Annnotation

Zielstellung

Das Ziel der Arbeit ist die Entwicklung einer Methode zur Klassifizierung von Dokumentinhalten um automatisch Metadaten zu erzeugen.

Machinelles Lernen (ML) ist eine offensichtliche Wahl für Klassifizierungsprobleme. Für die Aufgabe der Bildklassifikation liefern künstliche Neuronalle Netze derzeit die besten Ergebnisse[**rodrigo_benenson_classification_2016**] Die Klasse der "regional Convolutional Neural Networks" (sieh z.B. [**girshick_region-based_2016**]) ist in der Lage auf Basis eines CNN Bildregionen zu klassifizieren und zu segmentieren.

Die besten Ergebnisse werden mit "supervised learning" Verfahren erreicht. Solche Verfahren profitieren ernorm von großen annotierten Datensätzen, welche für die hier genannte Problemstellung nur begrenzt vorliegen. Um dieses Problem zu Lösen müssen in dieser Arbeit mehrere Ansätze kombiniert werden. CNN können sehr generische Bildrepräsentationen liefern [**razavian_cnn_2014**]. Deswegen können bereits trainierte NN "zweckentfremdet" werden. Deshalb soll untersucht werden welche anderen Datensets für das Training verwendet werden können. Zum anderen ist es möglich NN-Klassifizierer auf großen Mengen unannotierten Daten zu trainieren [**le_building_2013**].

Für einen effektiven Einsatz von ML müssen auch Werkzeuge programmiert werden um die vorhandenen Daten zu visuallisieren und (semi)manuellel zu klassifizieren. Ohne diese Vorarbeit ist eine Evaluierung der Ergebnisse nur schwer möglich.

1. Beschreibung des Umfangs der Digitalen Sammlung
2. Theorieteil über Neuronale Netze
3. Erstellung von Evaluierungsdaten
4. Auswahl von geeigneten Architekturen für NN
5. Parallelisierung der Datenverarbeitung
6. Evaluierung und Visuallisierung der Ergebnisse