Evaluation KI-basierter Lösungen zur Extraktion von Informationen aus Diagrammen

Mirko Sprcic
Technische Hochschule Mittelhessen
Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
mirko.sprcic@mnd.thm.de

Prof. Dr. Harald Ritz
Technische Hochschule Mittelhessen
Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
harald.ritz@mni.thm.de

Prof. Dr. Frank Kammer
Technische Hochschule Mittelhessen
Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
frank.kammer@mni.thm.de

Kategorie

Masterarbeit

Schlüsselwörter

Informationsextraktion, Künstliche Intelligenz, Deep-Learning, Computer Vision, Convolutional Neural Networks, Vision Transformer, Objekterkennung, Linienextraktion, Optical Character Recognition, YOLO, PyTorch, HuggingFace

Zusammenfassung

Neben der Informationsextraktion (IE) aus unstrukturierten Daten im Bereich natürlicher Sprache werden zunehmend Inhalte aus Bildern verarbeitet. Diese werden unter anderem zur Auswertung von Unternehmenskennzahlen oder zur Verbesserung der Zugänglichkeit von Inhalten in Bildern für Menschen mit Seheinschränkungen verwendet.

Ein weiterer Anwendungsfall existiert im Hochschulkontext. Abgaben von Studierenden zu Fallstudien oder Übungen bestehen häufig aus Bildern, die anschließend von Dozenten oder wissenschaftlichen Mitarbeitenden manuell kontrolliert und bewertet werden. Sind große Mengen an Bildern zu analysieren, so ist dieser Prozess für einzelne Personen sehr zeitaufwendig, weshalb Systeme zu entwickeln sind, die Inhalte in Bildern automatisiert extrahieren und weiterführenden Systemen zur Verfügung stellen.

Hierfür wird im Rahmen dieser Arbeit zunächst eine Pipeline zur IE aus Bildern konzipiert, die auf bisherigen Ansätzen und spezifischen Anforderungen des betrachteten Anwendungsfalls basiert. Im Rahmen der Pipeline werden Deep-Learning (DL)-Modelle eingesetzt, die für Aufgaben wie Bildklassifizierung oder Objekterkennung optimiert sind. Die Auswahl des verwendeten Modells für den jeweiligen Prozessschritt erfolgt durch eine Evaluation bestehender Modelle. Abschließend wird die erarbeitete Pipeline mit den evaluierten Modellen als Prototyp implementiert.

Der betrachtete Anwendungsfall besteht aus Abgaben zu Fallstudien des Wirtschaftsinformatik-Moduls Business Intelligence der Technischen Hochschule Mittelhessen. Die Abgaben umfassen eine Vielzahl an Bildschirmaufnahmen der SAP Business Warehouse Umgebung, darunter Tabellen, Transformationstabellen, Datenflussdiagramme, Eingabemasken und Bilder mit Mischformen der genannten Typen. Innerhalb der Bilder sind sowohl grafische als auch textliche Inhalte zu extrahieren und in Beziehung zueinander zu setzen.

Die entwickelte Pipeline besteht aus fünf Prozessschritten. Bilder werden zunächst klassifiziert, da sich die nachfolgenden Schritte in Abhängigkeit vom jeweiligen Typ unterscheiden. Handelt es sich bei dem zu verarbeitenden Bild um eine Mischform mehrerer Typen, wird das Bild zunächst in seine einzelnen Bestandteile zerlegt. Anschließend folgt die diagrammtypspezifische Extraktion zur Lokalisierung grafischer Elemente, wie Zeilen und Spalten in Tabellen oder Objekten in Datenflussdiagrammen. Hierfür werden überwiegend DL-Modelle verwendet, ergänzt durch regelbasierte Methoden zur Linienextraktion. Abschließend erfolgt die Erkennung von Texten mittels Optical Character Recognition (OCR) sowie die Zusammenführung der extrahierten Inhalte in einer einheitlichen JSON-Struktur.

Für die Klassifikation und Objekterkennung werden Modelle der Anbieter PyTorch, HuggingFace und Ultralytics evaluiert. Bei den Modellen handelt es sich um vortrainierte Modelle, die bereits mit großen Datenmengen trainiert wurden und dadurch in der Lage sind, relevante Merkmale aus Bildern zu extrahieren. Die Modelle werden mit Daten des Anwendungsfalls abgestimmt, um sie für die jeweilige Aufgabe nutzbar zu machen. Im Rahmen der Evaluation werden Modelle unterschiedlicher Architekturen und Größen betrachtet.

Neben den etablierten Convolutional Neural Networks (CNNs) werden Transformer-basierte Architekturen eingesetzt, die bereits im Bereich der Verarbeitung natürlicher Sprache erfolgreich verwendet werden.

Im Rahmen der Klassifikation erzielt der Großteil der Modelle einen gemittelten F1-Score von über 95 %. Unterschiede zeigen sich insbesondere in der Modellgröße und der erforderlichen Trainingszeit. Hier überzeugen besonders die Modelle RegNet, GoogLeNet und SqueezeNet.

Bei der Objekterkennung erzielen die von Ultralytics bereitgestellten YOLO-Modelle über sämtliche Aufgaben hinweg sehr gute Ergebnisse, mit Mean Average Precision (mAP)@50-Werten von über 98 % und mAP@50–95-Werten von über 80 %. Transformer-basierte DETR-Modelle liefern bei klar voneinander abgegrenzten Inhalten vergleichbare Ergebnisse, schneiden jedoch bei der Erkennung kleiner oder eng beieinanderliegender Objekte schwächer ab. Hier überzeugen hingegen CNN-Modelle durch ihre Fähigkeit, lokale Merkmale präziser zu erkennen.

Im Bereich OCR werden die Tools Tesseract, EasyOCR und PaddleOCR miteinander verglichen. Unterschiede zeigen sich bei der Verwendung mehrerer Sprachen. Während Tesseract und EasyOCR die Konfiguration mehrerer Sprachen gleichzeitig erlauben, unterstützt PaddleOCR ausschließlich die Nutzung einer Sprache pro Ausführung. Bei der Erkennungsqualität liefern EasyOCR und PaddleOCR mit einer Character Error Rate von 8,5 % bzw. 10,5 % ähnliche Ergebnisse, während Tesseract mit 30 % deutlich mehr Fehler erzeugt. Häufige Fehlerquellen treten bei der Erkennung ähnlich aussehender Zeichen auf, insbesondere bei Zahlen und Buchstaben.

Für die abschließende Zusammenführung der extrahierten Informationen werden regelbasierte Methoden eingesetzt. Dabei werden Abstände zwischen erkannten Objekten berechnet und mit definierten Toleranzen abgeglichen. Für die betrachteten Diagrammtypen funktioniert dieser Ansatz zuverlässig, allerdings ist bei der Einbindung weiterer Diagrammtypen zusätzlicher Entwicklungsaufwand erforderlich, da diese jeweils über eigene Regeln zur Zusammenführung verfügen.

Insgesamt zeigt die Arbeit, dass eine automatisierte Extraktion von Inhalten aus unterschiedlichen Diagrammtypen durch die Kombination vortrainierter DL-Modelle und regelbasierter Methoden erfolgreich umsetzbar ist. Der Einsatz regelbasierter Methoden zur Zusammenführung der Daten erfordert jedoch Codeanpassungen. Die Integration von Large Language Models (LLMs) ermöglicht eine Reduzierung dieser Methoden und der damit verbundenen Anpassungen bei Erweiterungen des Systems. LLMs könnten auf Basis der extrahierten Inhalte sowie eines diagramm-typspezifischen Prompts, der die gewünschte Zusammenführung beschreibt, die finalen Strukturen erzeugen. Der Aufwand zur Erweiterung um neue Diagrammtypen würde sich dadurch auf das Training geeigneter DL-Modelle zur Objekterkennung und die Erstellung eines Prompts, der das Verhalten des Sprachmodells steuert, reduzieren. Voraussetzung hierfür ist jedoch der Einsatz leistungsfähiger LLMs mit entsprechender Rechenkapazität, da kleinere Modelle aktuell noch nicht über die erforderlichen Fähigkeiten im Bereich der Sprachverarbeitung verfügen.

Literatur

Charles, P. K. et al. (2012): A Review on the Various Techniques used for Optical Character Recognition. In: International Journal of Engineering Research and Applications 2 (1), S. 659–662.

Choi, J. et al. (2019): Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization. In: Computer Graphics Forum 38 (3), S. 249–260. DOI: 10.1111/cgf.13686.

Cowie, J.; Lehnert, W. (1996): Information extraction. In: Commun. ACM 39 (1), S. 80–91. DOI: 10.1145/234173.234209.

Davila, K. et al. (2021): ICPR 2020 - Competition on Harvesting Raw Tables from Infographics. In: Pattern Recognition. ICPR International Workshops and Challenges 12668, S. 361–380. DOI: 10.1007/978-3-030-68793-9_27.

Dosovitskiy, A. et al. (2020): An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Online verfügbar unter http://arxiv.org/pdf/2010.11929v2.

Liu, F. et al. (2023): DePlot: One-shot visual language reasoning by plot-to-table translation. In: Findings of the Association for Computational Linguistics: ACL 2023, S. 10381–10399. DOI: 10.18653/v1/2023.findings-acl.660.

Luo, J. et al. (2021): ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), S. 1916–1924. DOI: 10.1109/WACV48630.2021.00196.

Maheshwari, K. et al. (2020): Convolutional Neural Networks: A Bottom-Up Approach. In: Deep Learning: De Gruyter, S. 21–50. DOI: 10.1515/9783110670905-002.

Masry, A. et al. (2023): UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. In: Proceedings of the 2023 Conference on

Empirical Methods in Natural Language Processing, S. 14662–14684. DOI: 10.18653/v1/2023.emnlp-main.906.

Mustafa, O. et al. (2023): ChartEye: A Deep Learning Framework for Chart Information Extraction. In: 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), S. 554– 561. DOI: 10.1109/DICTA60407.2023.00082.

Wang, H. et al. (2023): Pre-Trained Language Models and Their Applications. In: Engineering 25, S. 51–65. DOI: 10.1016/j.eng.2022.04.024.

Zhang, H. et al. (2013): Text extraction from natural scene image: A survey. In: Neurocomputing 122, S. 310–323. DOI: 10.1016/j.neucom.2013.05.037.