Der vorliegende Artikel beschreibt die Entstehung und Weiterentwicklung der Bildähnlichkeitssuche an der Bayerischen Staatsbibliothek. Bereits 2013 wurden Verfahren des maschinellen Lernens eingesetzt, um ein System für die Bildsuche im Gesamtbestand zu ermöglichen. Angetrieben von den enormen technischen Entwicklungen der letzten Jahre im Bereich Künstliche Intelligenz (KI) wird aktuell an einer neuen Version der Bildsuche gearbeitet. Die für 2023 geplante neue Bildähnlichkeitssuche verwendet dabei aktuelle Technologien aus dem Bereich des maschinellen Lernens. Es kommen damit erstmals tiefe neuronale Netze zum Einsatz, die in Verbindung mit einer leistungsfähigen Suchmaschine deutlich verbesserte, eindrucksvolle Ergebnisse bei der Suche nach ähnlichen Bildern hervorbringen. Die gesammelten Erfahrungen und Herausforderungen bei der Umsetzung des neuen Systems werden in diesem Beitrag präsentiert.

This article describes the establishment and further development of image similarity searches at the Bayerische Staatsbibliothek. Back in 2013, machine learning methods were already being used as the basis for a system which could search for images in the library's holdings. Prompted by the major technical advances in artificial intelligence (AI) in recent years, work is currently underway on the development of a new version of the image search technology. The new image similarity search service is planned for 2023 and is based on the latest machine learning technologies. This is the first time that deep neural networks have been used. In combination with a powerful search engine, these produce impressive, significantly improved results in searches for similar images. This paper presents the experience accrued to date and challenges in implementing the new system.

MARKUS BRANTL, STEFAN SCHWETER

Neue Wege der Bildsuche an der Bayerischen Staatsbibliothek

Seit 2013 bietet das Münchener Digitalisierungszentrum (MDZ) der Bayerischen Staatsbibliothek (BSB) mit der Webanwendung »Bildähnlichkeitssuche« eine KI-unterstützte Bilddatenbank zur Suche nach ähnlichen Bildern bzw. Bildsegmenten an. Über einen Upload-Mechanismus können eigene Bilder zur vergleichenden Suche geladen und eine Suche über den Gesamtbestand ausgeführt werden. 57 Millionen automatisch aus den 383 Millionen digitalisierten Buchseiten extrahierte Bilder bzw. Bildsegmente können so nach den Ähnlichkeitsmerkmalen Farb- und Kanteninformationen durchsucht werden.[1] Die digitalen Bilder stammen aus dem Gesamtbestand der »Digitalen Sammlungen«, die seit 1997 an der Bayerischen Staatsbibliothek im Rahmen der Retrodigitalisierung und der Kooperation mit Google Books (seit 2007) produziert wurden und werden. Auf einem Zeitstrahl betrachtet, handelt es sich um Werke aus einem Zeitraum vom 7. Jahrhundert v. Chr. bis ins 20. Jahrhundert: also vom griechischen Papyrusfragment über mittelalterliche Handschriften mit illuminierten Miniaturen, Holzschnitte der frühen Neuzeit bis hin zu feinen Stahlstichen und gerasterten Fotografien z. B. in den Zeitungen des 20. Jahrhunderts. Die auf den einzelnen Buchseiten vorhandenen Bilder bzw. Bildsegmente gilt es zu erkennen, extrahieren, klassifizieren, speichern und mit einer entsprechenden Suchfunktionalität bereitzustellen.

Die erste, jetzt noch im Einsatz befindliche Version der Bildähnlichkeitssuche wurde damals in Zusammenarbeit mit dem Fraunhofer Heinrich-Hertz-Institut (HHI), das bereits Erfahrung mit der performanten Bildanalyse von mehreren Millionen Bildern hatte, entwickelt. Als Basis diente eine für die Detektion von Plagiaten moderner Digitalfotografien im Internet entwickelte Softwarelösung des HHI. Im Rahmen des gemeinsamen Entwicklungsprojekts von HHI und BSB wurden damals bereits Verfahren des maschinellen Lernens z. B. bei der »Text-Bild-Trennung« oder bei der Optimierung der Bilderkennungsergebnisse eingesetzt.[2]

Die Digitalen Sammlungen[3] der BSB wurden im April 2021 als national größtes Angebot an Digitalisaten des schriftlichen Kulturerbes mit fast 3 Millionen digitalisierten Werken einer kompletten technischen Erneuerung unterzogen. Fast alle Digitalisate (98 %) sind nun über die offenen Image- und Presentation-API (Application Programming Interface) des International Image Interoperability Frameworks (IIIF) und mit dem IIIF-konformen Mirador-Viewer[4] in einem komfortablen Zugang und einer neuen User Experience (UX) weltweit zugänglich. Bereichert wird die neue WWW-Applikation der Digitalen Sammlungen durch die vollständige Bereitstellung und Durchsuchbarkeit aller maschinenlesbaren Volltexte beginnend ab dem 16. Jahrhundert. Durch die ständig optimierte Google-OCR, die Be-

standteil der Kooperation mit Google Books ist und iterativ an die BSB bzw. das MDZ geliefert wird, stehen Volltexte mit Layouterkennung für über 250 Sprachen und mehr als 30 Schriften zur Nutzung zur Verfügung. Diese werden mit einem zentralen Solr-Index verarbeitet und bereitgestellt.[5]

Die OCR stößt aber im Bereich der mittelalterlichen Handschriften und Spezialbestände noch an die Grenzen der maschinellen Lesbarkeit. Zwar sind alle Werke der BSB durch umfangreiche, qualifizierte und online durchsuchbare bibliografische Metadaten indexiert und recherchierbar, aber die Beschreibung der Struktur einer mittelalterlichen Handschrift, also z. B. die Erfassung von Überschriften, Kapitelnummern, Illustrationen ist derzeit noch eine aufwendige und teure Handarbeit. Somit bliebe ohne eine Bildähnlichkeitssuche der Zugang für die Nutzer*innen zu den bildlichen Inhalten eines Buches solange verschlossen, bis das Werk am Bildschirm Seite für Seite oder in einer Thumbnail-Ansicht durchblättert wird.

Mit der neuen Internetbereitstellung der Digitalen Sammlungen, der laufend wachsenden Datenmenge durch BSB-eigene und die Google-Produktion sowie die technischen Innovationen wuchs der Anspruch an die Leistungsfähigkeit der in die Jahre gekommenen technischen Lösung für die Bildähnlichkeitssuche. Technisch war 2021 auch die Hardware für die Verarbeitung der Bilder (Extraktion, Indexerstellung etc.), ein 37 Server umfassendes Linux-Cluster, am Ende ihres Lebenszyklus angekommen. Mit den ersten Überlegungen für den Austausch der Hardware wurde rasch die Zielrichtung des technischen Neuansatzes deutlich: das gesamte Verfahren sollte ausschließlich mit Verfahren des maschinellen Lernens auf einer neuen, technisch performanteren GPU-basierten (Graphical Processing Unit) Server-Infrastruktur stattfinden.

Die folgende Darstellung beschreibt als Werkstattbericht den Weg zur neuen Bildsuche der Bayerischen Staatsbibliothek, die zunächst 2023 im Rahmen eines Pilotprojekts für das Kulturportal des Freistaats Bayern »bavarikon« live gehen soll und anschließend auch Anwendung auf den Gesamtbestand der Digitalen Sammlungen mit der Integration der Bildsuche in die IIIF-Welt (Mirador) finden soll.

Die Herausforderungen der alten und der neuen Bildsuche

Bevor es in die technischen Details geht, seien hier nochmals die Herausforderungen des alten und neuen Projekts kurz zusammengefasst:

1. Die Datenmenge, die es zu prozessieren gilt: 383 Millionen Bildseiten sind auf Abbildungen zu durchsuchen und diese zu extrahieren. Im Fall der alten Bildsuche waren dies dann 57 Millionen Bilder bzw. Bildsegmente, wobei ca. bis zu 20 % der Bilder beziehungsweise Bildsegmente aufgrund der Probleme mit der vorgeschalteten Text-Bild-Trennung nicht genutzt werden konnten.
2. Die materialimmanenten Herausforderungen bei der Analyse der Seiten, wie etwa kunstvolle Miniaturen auf mittelalterlichem Pergament, Holzschnitte auf Büttenpapier, kolorierte Kupferstiche in unterschiedlichsten Größen, verwoben mit Textelementen, filigrane Stahlstiche auf Seidenpapier und nicht zuletzt grob gerasterte Fotografien auf stark vergilbtem, billigem und brüchigem Papier von Büchern und Zeitungen der Neuzeit. Dabei sind Artefakte in den Seiten selbst möglichst von der Erkennung zu eliminieren, wie etwa Stockflecken, Risse, Eselsohren in Buchseiten.
3. Die automatisierte Erkennung von Bildern und gegebenenfalls deren Extraktion aus einer Seite mit Textinformation bzw. von einzelnen Bildern aus einer Buchseite, die mehr als eine Illustration enthält. Es sollten nach Möglichkeit natürlich alle Bilder einer Seite einzeln erkannt werden und auf dem Image lokalisiert und extrahiert werden.
4. Extraktion und Beschreibung der visuellen Merkmale der Bilder. Bei der bisherigen Anwendung waren dies spezifische Kanten- und Farbinformationen, die in einem Deskriptor pro Bild erfasst wurden.

Wie funktioniert die neue Bildähnlichkeitssuche

Die neue Bildähnlichkeitssuche verwendet aktuelle Technologien aus dem Bereich des maschinellen Lernens. Beim maschinellen Lernen lernt ein künstliches System – in unserem Fall ein künstliches neuronales Netz – anhand von Beispielen und kann dieses angeeignete Wissen auf neue, vorher noch nicht gesehene Beispiele übertragen. Künstliche neuronale Netze sind bis zu einem gewissen Grad dem Aufbau des menschlichen Gehirns nachempfunden. Stark abstrahiert dargestellt, besteht dieses neuronale Netz aus Millionen von Neuronen (= Nervenzellen), die speziell angeordnet und verknüpft sind. Wie im menschlichen Gehirn sind die Verbindungen zwischen den Neuronen veränderbar und passen sich durch Training an äußere Reize an, sodass das Netzwerk lernen kann, verschiedene Aufgaben wie Objekterkennung oder auch das Übersetzen von Sprache zu lösen.

Um das künstliche neuronale Netz für die Bildähnlichkeitssuche zu trainieren, muss zunächst ein Trainingskorpus erstellt werden. Dafür werden Bildsegmente von Digitalisaten per Hand annotiert. Der Begriff Annotation bezieht sich dabei sowohl auf die Position (Koordinaten) als auch auf die gefundene Kategorie (wie Illustration, Illumination oder Person) des Bildsegments. Die per Hand annotierten Bildsegmente werden anschließend dazu verwendet, einen Klassifikator zu trainieren. Der Klassifikator ist in diesem Fall das künstliche neuronale Netz, das nach dem Training Merkmale

in Bildern erkennen und Bildsegmente anhand dieser Merkmale in verschiedene Klassen einteilen kann. Später extrahiert dieser Klassifikator Bildsegmente automatisiert aus einem gewünschten Bildbestand. Aus diesen extrahierten Bildsegmenten werden durch ein weiteres künstliches neuronales Netz Merkmalsvektoren erzeugt, die dann in eine Vektorsuchmaschine indexiert werden. Dadurch sind schnellere und bessere Bildähnlichkeitssuchen im Vergleich zur alten Bildähnlichkeitssuche möglich. Im Folgenden werden die Schritte genauer erklärt, die zur Implementierung der neuen Bildähnlichkeitssuche notwendig waren. Dabei wird auf die Herausforderungen und gewonnenen Erkenntnisse eingegangen, die bei der Implementierung zu bewältigen waren.

Manuelle Annotation von Bildsegmenten

Im Fall der neuen Bildähnlichkeitssuche setzen sich die Beispiele aus annotierten Bildsegmenten zusammen. Diese Bildsegmente können Illustrationen, Illuminationen, Personen oder andere Objekte sein. Beim sog. Annotationsprozess müssen viele Buchseiten manuell durchgesehen werden. Dafür wurde die Software »Prodigy«[6] verwendet. Prodigy ermöglicht die Annotation von Bildern mit einer komfortablen webbasierten Benutzeroberfläche, in der u. a. Tastaturkürzel für das schnelle Kategorisieren von Bildsegmenten verwendet werden können. In einem Testlauf konnten so ca. 100 Bilder pro Stunde annotiert werden. Prodigy ermöglichte somit die schnelle und effiziente Umsetzung des Annotationsprozesses.

Nach dem durchgeführten Annotationstestlauf kamen diverse Fragen bezüglich einer »guten« Annotation von Bildsegmenten auf. So stellte sich beispielsweise die Frage, ob größere Bilder, wie Karten oder zusammengesetzte Abbildungen, nicht doch in kleinere Bildsegmente unterteilt werden sollten und ob generell verschachtelte Annotationen sinnvoll wären. Des Weiteren stellte sich die Frage, ob beispielsweise Tabellen ebenfalls als Bildsegmente behandelt werden sollten. Um diese Fragen abschließend zu klären, muss ein künstliches System aus eben diesen Bildsegmenten gelernt werden, damit die Erkennungsleistung auf ungesehenen Bildern bewertet werden kann.

Gängige Verfahren bei der Objekterkennung

Um generell ein System für Objekterkennung zu trainieren, muss eine Vielzahl an annotierten Bildern vorhanden sein. Eine gängige Datenquelle dafür ist »ImageNet«,[7] welches für Forschungszwecke eingesetzt wird. ImageNet wurde 2009 erstmals veröffentlicht und umfasst mehr als 14 Millionen Bilder. Jedes Bild in ImageNet wurde per Hand annotiert und darauf erkannte Objekte in eine von ca. 20.000 festgelegten Kategorien eingeteilt. Seit 2010 wird vom ImageNet-Projekt jährlich ein Wettbewerb veranstaltet, um das beste System für die korrekte Klassifizierung und Erkennung von Objekten zu finden. 2016 wurde das »ResNet«-Modell[8] vorgestellt, das die Elemente auf Bildern genauso gut erkennen kann wie ein Mensch und daher als Meilenstein bei der Bilderkennung mit Methoden des maschinellen Lernens betrachtet werden kann.

2017 veröffentlichten Forscher*innen von Google Brain die sogenannte Transformer-Architektur.[9] Diese Architektur kann man sich als eine Art Methode vorstellen, die eine Folge von Zeichen in andere Folgen von Zeichen übersetzen kann. Die Architektur wurde fortan für z. B. maschinelle Übersetzung (wie bei Google Translate[10]) verwendet, um effizient Systeme auf großen Datenmengen zu trainieren. Neben der maschinellen Übersetzung können mithilfe von Transformern auch Sprachmodelle[11] trainiert werden, die die Grundlage für aktuelle Frage-Antwort-Systeme (»Question Answering«) bilden oder auch für das automatische Zusammenfassen von Texten (»Summarization«) verwendet werden.[12] Transformer-basierte Modelle sind aktueller Stand der Technik und kommen ebenfalls im Bereich Bilderkennung zum Einsatz. Architekturen wie »Swin Transformer«[13] oder »DeiT«[14] (kurz für »Training data-efficient Image Transformer«) gehören aktuell zu den besten Architekturen für Objekterkennung und basieren auf Transformern.

Bevor Transformer-basierte Architekturen für die Bilderkennung zum Einsatz kamen, wurden meistens sogenannte »Convolutional Neural Networks«[15] (»CNN« oder »ConvNet«) verwendet. Die Ideen für diese spezielle Architekturen reichen bis in die 1950er-Jahre zurück und wurden Ende der 1980er-Jahre schon erfolgreich für das Erkennen von handgeschriebenen Postleitzahlen eingesetzt.[16] 1998 wurde das sog. »LeNet-5«-Modell[17] verwendet, um handgeschriebene Ziffern auf Bankschecks automatisiert zu erkennen. 2012 gewannt das »AlexNet«-Modell[18], welches ebenfalls auf CNNs basiert und auf GPUs trainiert wurde, den vorher erwähnten ImageNet-Wettbewerb.

Training eines Klassifikators zur automatischen Erkennung von Bildsegmenten

Für das Lernen eines Systems auf den annotierten Bildsegmenten für die neue Bildähnlichkeitssuche wurde eine von Meta AI (ehemals Facebook AI) entwickelte Open Source Software namens »Detectron2«[19] verwendet. Diese Software lernt anhand der annotierten Beispiele Bildsegmente eigenständig zu erkennen und zu klassifizieren. Dieses Verfahren wird auch als überwachtes Lernen bezeichnet. Ziel ist es, dass der Klassifikator anhand der annotierten Bildsegmente eigenständig wichtige Merkmale findet, die charakteristisch für zum Beispiel Illustrationen wie Wappen oder Münzen sind. Im Laufe des Trainingsprozesses des Klassifikators hat sich gezeigt, dass nicht nur Bilder mit vorhandenen Bildsegmenten verwendet werden sollten, sondern dass ebenfalls Bilder mit aufgenommen werden sollten, auf

denen überhaupt keine Annotationen zu finden sind. Diese Bilder ohne Annotationen dienen dann als Negativbeispiele im Trainingsprozess und verbesserten die spätere Erkennungsleistung enorm. Für den ersten Prototypen der Bildähnlichkeitssuche wurden ca. 1.000 Bilder mit insgesamt etwa 100 annotierten Bildsegmenten eingesetzt. Das Training des Klassifikators erfolgte dabei auf spezieller Hardware (Grafikkarten, sog. GPUs), da somit der Trainingsprozess enorm beschleunigt werden kann. So dauerte das Training rund zwei Stunden.

Für das Trainieren des Klassifikators mittels Detectron2-Bibliothek können sowohl CNN-basierte als auch Transformer-basierte Modelle verwendet werden. In Vorabexperimenten konnte festgestellt werden, dass die Menge an annotierten Bildsegmenten nicht ausreicht, um Transformer-basierte Modelle zu trainieren. Die Ergebnisse bei der Erkennung von Bildsegmenten ist deutlich schlechter als mit CNN-basierten Modellen. Daher basiert der trainierte Klassifikator auf CNNs, der mittels Detectron2-Bibliothek trainiert wird. Des Weiteren ist festzustellen, dass CNN-basierte Modelle deutlich schneller trainiert werden können als Transformer-basierte Architekturen. Auch das Anwenden eines trainierten Modells auf Bilder ist mit CNN-basierten Modellen in der Regel schneller als Transformer-basierte Lösungen.

Evaluation

Bei überwachtem Lernen wird üblicherweise das gelernte System nach dem Training getestet bzw. evaluiert. Dazu werden Testdaten verwendet, die unter keinen Umständen Teil des eigentlichen Trainings waren. Nur so kann wirklich überprüft werden, wie gut das System generalisiert. Für die Evaluationsphase wurden händisch 50 Bilder ausgewählt, um zu überprüfen, ob gewünschte Segmente von dem System erkannt werden. In der ersten Evaluationsphase konnten folgende Erkenntnisse gewonnen werden: Wurden im Annotationsschritt beispielsweise Tabellen als Bildsegment markiert, so hatte dies gravierende Auswirkungen auf die Erkennungsleistung. Da in den annotierten Tabellen meistens Zahlen vorhanden waren, konnte beobachtet werden, dass nun Seitenzahlen auf Buchseiten fälschlicherweise als Bildsegment erkannt wurden. Wenn Abbildungen wie große Karten als ein Bildsegment annotiert wurden, wurden farbige und verzierte Buchdeckel ebenfalls als Bildsegment erkannt.

Nachdem diese Probleme während der Evaluationsphase aufgetaucht waren, fand eine Nachannotation der bereits vorhandenen Trainingsdaten statt. Dabei wurden beispielsweise Tabellen oder reiner Text, der fälschlicherweise annotiert wurde, entfernt. Des Weiteren wurden Karten in kleinere Bildsegmente unterteilt. So kann beispielsweise eine Legende als eigenständiges Bildsegment betrachtet werden, oder entsprechende Logos auf der Karte. Nach dieser manuellen Nachkorrektur wurde wieder ein neues System trainiert und evaluiert. Es konnte nun festgestellt werden, dass es viel weniger falsch erkannte Bildsegmente gab. Eine weitere Erkenntnis bei der manuellen Nachkorrektur war, dass oft nur wenige Trainingsbeispiele ausreichen, um später eine zuverlässige Erkennung zu gewährleisten. So reicht es beispielsweise, dass eine einstellige Anzahl unterschiedlicher Münzen oder Siegel annotiert wird, um später zuverlässig noch nicht gesehene Münzen oder Siegel zu erkennen. Dies gilt allerdings auch für Fehlannotationen: oft reichen nur wenige annotierte Bildsegmente wie Tabellen aus, um unerwünschte Nebeneffekte, wie erkannte Bildsegmente, die nur aus Seitenzahlen bestehen, zu erzielen.

Active-Learning-Prinzip

Für die erste Version des Klassifikators wurden wie oben erwähnt ca. 1.000 Bilder mit rund 100 annotierten Bildsegmenten verwendet, die in einem Nachannotationsprozess korrigiert wurden. Für eine weitere Version des Klassifikators wurde eine Technik namens »Active-Learning«[20] verwendet, um weitere Trainingsdaten zu erzeugen und die Erkennungsleistung des Klassifikators zu verbessern. Dabei werden dem Klassifikator nicht bekannte Bilder gezeigt, auf denen Bildsegmente erkannt werden sollen. Die Erkennungsleistung wird manuell überprüft, und es werden bei Bedarf Korrekturen an den erkannten Bildsegmenten vorgenommen, wie beispielsweise die Größenanpassung des erkannten Bildausschnittes, Löschen von Fehlannotationen oder das Unterteilen einer großen Abbildung in kleinere. Auf diesen nach-annotierten Bildsegmenten wird nun erneut ein Klassifikator trainiert, der im Vergleich zur ersten Version nun aber deutlich mehr Trainingsdaten als Grundlage hat. Dieser Prozess kann beliebig oft wiederholt werden, weshalb das Active-Learning-Prinzip als eine Art Kreislauf verstanden werden kann. Da durch diesen Kreislauf immer weniger nachkorrigiert werden muss, ist das Active-Learning-Prinzip sehr effizient. Für die aktuelle Version des Klassifikators werden ca. 4.300 Bilder mit rund 700 annotierten Bildsegmenten verwendet. Diese Anzahl an Trainingsdaten ermöglicht sehr gute Erkennungsraten in unseren Experimenten.

Erzeugung von Merkmalsvektoren für Bildsegmente

Um eine Suche nach ähnlichen Bildsegmenten bzw. eine Bildähnlichkeitssuche zu ermöglichen, muss für jedes Bildsegment ein sog. Merkmalsvektor erzeugt werden. Dieser Merkmalsvektor ist ein höher-dimensioneller Vektor, der später in einem Vektorraum dargestellt werden kann. In unseren Experimenten kamen verschiedene Dimensionsgrößen (192, 512 und 2.048) zum Einsatz. Erzeugt werden diese Merkmalsvektoren u. a. von künstlichen neuronalen Netzen wie ResNet-50 oder DeiT, die auf der Bilddatenbank ImageNet trainiert wurden.

Für den Prototypen der neuen Bildähnlichkeitssuche kamen sowohl Transformer-basierte Modelle als auch CNN-basierte Modelle zur Erstellung der Merkmalsvektoren zum Einsatz. Transformer-basierte Modelle können für diesen Zweck verwendet werden (im Gegensatz zum vorher erwähnten Klassifikator-Training), da das künstliche neuronale Netz zum Extrahieren der Merkmalsvektoren nicht neu trainiert werden muss: es wird ein bereits trainiertes Modell verwendet.

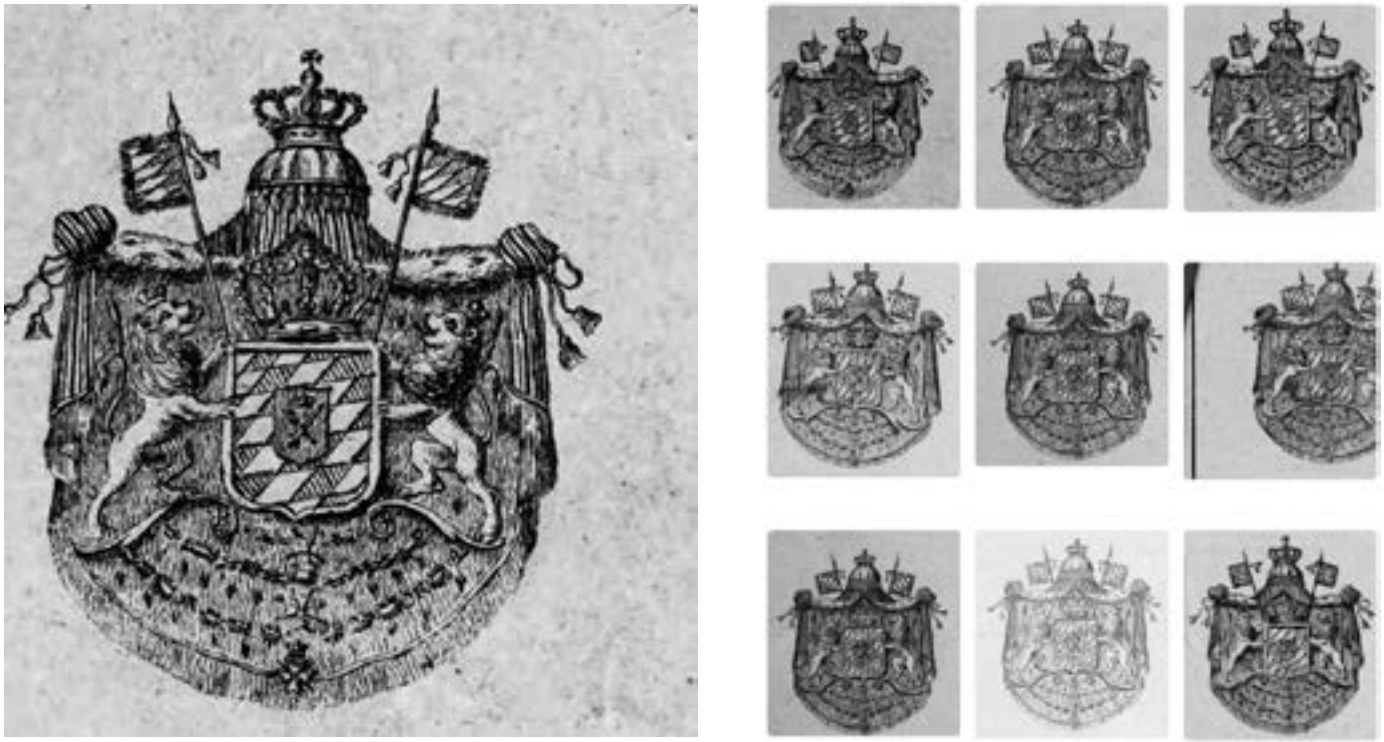
Es konnte beobachtet werden, dass eine Dimensionsgröße von 192 bereits ausreicht, um sehr gute Ergebnisse für die Bildähnlichkeitssuche zu erreichen. Der Unterschied zu Vektoren, die mit einer Dimensionsgröße von 2.048 erstellt werden, ist nur marginal. Das Verwenden von Vektoren, die 192 statt 2.048 Dimensionen besitzen, bringt weitere Vorteile mit sich: da die Vektoren kleiner sind, wird weniger Speicherplatz benötigt. Des Weiteren ist die Suche nach ähnlichen Vektoren mit kleiner Dimensionsgröße wesentlich schneller. Zusammenfassend kann man also sagen, dass diese Lösung wesentlich effizienter ist und weniger Ressourcen benötigt bei annähernd gleicher Erkennungsleistung.

Vektorsuchmaschine

Bevor allerdings eine Bildähnlichkeitssuche durchgeführt werden kann, müssen die erzeugten Vektoren effizient gespeichert werden. Dafür wird die Open Source Suchplattform »Solr«[21] verwendet. Solr kann ab Version 9[22] auch Vektoren indexieren und bietet zusätzlich eine Funktion, um Vektorähnlichkeitssuchen auszuführen. Dadurch, dass die Bildsegmente nun als Vektoren gespeichert sind, können durch Kosinusähnlichkeit ähnliche Bilder/Bildsegmente effizient gesucht und zurückgegeben werden. Eine Suchabfrage in Solr bewegt sich im Millisekundenbereich und benötigt keine spezielle Hardware, wie sie beispielsweise beim Trainieren des Klassifikators zum Einsatz kam.

Beispielsuchen nach Personen und Wappen sind in den Abbildungen 1 und 2 zu sehen. Für die Bildähnlichkeitssuche nach einer Person ist in Abbildung 1 links das Ausgangsbild für die Suche zu sehen. Abbildung 1 rechts zeigt dann eine Übersicht an ähnlichen Personen, die gefunden wurden. Für die Bildähnlichkeitssuche nach einem Wappen ist in Abbildung 2 links das Ausgangsbild für die Suche zu sehen und in Abbildung 2

1 Ausgangsbild für die Suche (li) und die Übersicht an ähnlichen Personen, die gefunden wurden (re)
Abb.: Bayerische Staatsbibliothek



2 Ausgangsbild für die Suche (li) und die Übersicht an ähnlichen Wappen (re)
Abb.: Bayerische Staatsbibliothek

rechts die Übersicht an ähnlichen Wappen.

Anhand der Beispielsuchen kann man sehen, dass die Erkennungsleistung sehr gut ist. Bei der Suche nach ähnlichen Personen werden auch Personen gefunden, bei denen beispielsweise die Kopfposition deutlich anders als das Eingangsbild ist, aber trotzdem eine Ähnlichkeit besteht. Bei der Suche nach ähnlichen Wappen ist festzustellen, dass die Hintergrundfarbe keine große Rolle zu spielen scheint: so werden auch ähnliche Wappen gefunden, die eine deutlich hellere Hintergrundfarbe haben als das Eingangsbild.

Änderungen im Vergleich zur alten Bildähnlichkeitssuche

Auf Technikseite bedeutet die neue Bildähnlichkeitssuche folgende Umstellungen:

Dadurch, dass ein Klassifikator mit manuell annotierten Bildern trainiert wird, entfällt die Umsetzung einer »Text-Bild-Trennung« komplett. Durch die manuell annotierten Trainingsdaten ist der Klassifikator sehr gut in der Lage, Bildsegmente von Texten zu unterscheiden, da sich diese Unterscheidung in den Trainingsdaten widerspiegelt. Auf den Einsatz von morphologischen Operationen kann daher bei der neuen Bildähnlichkeitssuche verzichtet werden. Dies bedeutet auch, dass der zu erfassende Bildbestand nicht mehr extra durchlaufen werden muss, wie es bei der alten Bildähnlichkeitssuche der Fall war, um die »Text-Bild-Trennung« technisch umzusetzen. Dies stellt einen erheblichen Effizienzgewinn dar.

Basis der alten Bildähnlichkeitssuche sind die unterschiedlichen visuellen Eigenschaften eines Bildes, seine spezifischen Farb- und Kanteninformationen. Diese Informationen mussten dazu manuell festgelegt und technisch umgesetzt werden. Diese Arbeit ist nun nicht mehr notwendig, da relevante Bildsegmente durch den trainierten Klassifikator sehr gut erkannt werden. Dadurch, dass die erkannten Bildsegmente in ein künstliches neuronales Netz (wie ResNet oder DeiT) gegeben werden, welches auf Millionen von Bildern trainiert wurde, können visuelle Eigenschaften viel besser erkannt werden. Die daraus resultierenden Merkmalsvektoren liefern deutlich bessere Ergebnisse als die Merkmalsvektoren der alten Bildähnlichkeitssuche.

Bei einer manuellen Evaluation der extrahierten Bildsegmente der alten Bildähnlichkeitssuche konnten folgende häufige Fehler beobachtet werden: In vielen Fällen wurde auch dort, wo nur Text ist, dieser Text fälschlicherweise als Bildsegment erkannt. Auch komplette oder fast komplette Buchseiten sowie zu eng ausgeschnittene Bildsegmente waren die häufigsten Fehlerquellen bei der alten Bildähnlichkeitssuche. Bei der Umsetzung der neuen Bildähnlichkeitssuche wurde daher das Augenmerk verstärkt auf diese Arten von Fehlern gelegt, um diese Fehlerquellen in Zukunft weitgehend zu vermeiden. Eine laufende Qualitätskontrolle der ex-

trahierten Bildsegmente ist daher unbedingt notwendig, besonders nach dem Training eines neuen Klassifikators.

Auf Anwenderseite bedeutet die Umsetzung der neuen Bildähnlichkeit folgende »Änderungen«:

Es können weiterhin eigene Bilder hochgeladen werden, oder eine URL zu einem bestehenden Bild angegeben werden. Auch Links zu Bildern bzw. Bildausschnitten können als Eingabe fungieren. Dadurch, dass künstliche neuronale Netze die Extraktion von Merkmalen vornehmen, fallen Filtermöglichkeiten wie beispielsweise das Farb- oder Kantenähnlichkeitsmaß weg. Es kann nun aber ein Mindestmaß für die Bildähnlichkeit als Filter verwendet werden.

Weitere Vorteile, die sich durch die neue Bildähnlichkeitssuche ergeben, sind folgende: Dadurch, dass ein Klassifikator trainiert werden kann, besteht die Möglichkeit, auch neue Klassen von Bildsegmenten zu annotieren und somit bessere Ergebnisse für die Ähnlichkeitssuche zu erzielen, so wäre z. B. das neu digitalisierte *stern*-Fotoarchiv der Bayerischen Staatsbibliothek bzw. Teile daraus ein Anwendungsfall für eine solche neue Klasse. Durch aktuelle Technologien wie Solr können Vektorähnlichkeitssuchen viel effizienter und schneller durchgeführt werden, ohne dass Spezialhardware zum Einsatz kommen muss.

Dadurch, dass die Suchplattform Solr für die Bildähnlichkeitssuche verwendet wird, ergeben sich viele neue Verbesserungen auf Anwenderseite: Suchergebnisse können nun durch sog. Facetten gefiltert werden. Eine Facette kann beispielsweise eine Datums- bzw. Zeitangabe wie ein gewünschter Zeitraum sein. Des Weiteren können die Suchergebnisse später durch eine Medienart-Facette auf zum Beispiel Zeitungen oder Büchern eingegrenzt werden. Auch eine Kombination aus mehreren Facetten beispielsweise aus einer Zeit-Facette und einer Sammlungs-Facette ist dank Solr effizient möglich.

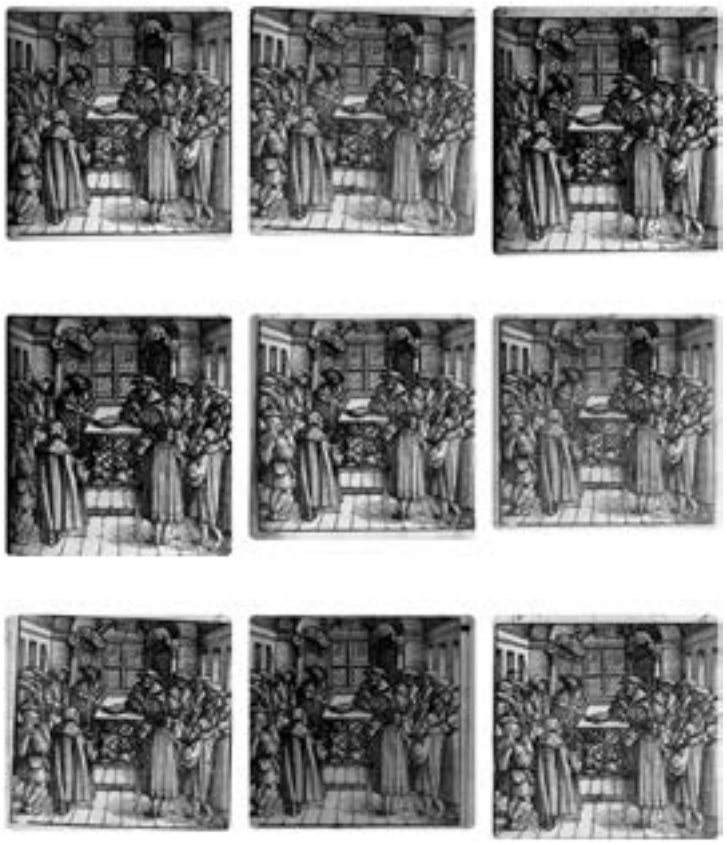
Synergieeffekte durch einen Klassifikations-Workflow lassen sich in anderen Bereichen ebenfalls erzielen: so können eigene Klassifikatoren trainiert werden, die beispielsweise bei der Qualitätskontrolle zum Einsatz kommen könnten. Konkrete Szenarien wären beispielsweise das Erkennen von Fehlern beim Scanvorgang, wie eingeknickte Seiten oder Fremdobjekte auf dem Scan. Darüber hinaus können eigene Klassifikatoren trainiert werden, um beispielsweise problematische Inhalte zu erkennen. Dies wird dadurch erreicht, dass der Klassifikator nur sehr wenige Bildsegmente benötigt.

Bildähnlichkeitssuchen: Neu gegen Alt

In einem weiteren Experiment ist die Erkennungsleistung der alten und neuen Bildähnlichkeitssuche un-

3 Ausgangsbild für die Ähnlichkeitssuche
Abb.: Bayerische Staatsbibliothek

4 Ergebnisse der alten Bildähnlichkeitssuche zu diesem Wappen
Abb.: Bayerische Staatsbibliothek



5 Ergebnisse mit der neuen Bildähnlichkeitssuche zu diesem Wappen
Abb.: Bayerische Staatsbibliothek

tersucht worden. So wurde beispielsweise nach einer ähnlichen Illustration gesucht, wie in Abbildung 3 zu sehen ist. Die Ergebnisse der alten Bildähnlichkeitssuche zu diesem Wappen sind in Abbildung 4 zu sehen. In Kontrast dazu sind die Ergebnisse mit der neuen Bildähnlichkeitssuche in Abbildung 5 dargestellt.

Es ist hier also festzustellen, dass die neue Bildähnlichkeitssuche wesentlich bessere Ergebnisse bzw. Bildsegmente findet als die alte Bildähnlichkeitssuche. Bei den Ergebnissen der alten Bildähnlichkeitssuche fällt auf, dass nur bedingt ähnliche Illustrationen in den Ergebnissen zu finden sind. Unter den ähnlichen Bildsegmenten sind auch ganze Buchdeckel oder ganze Seiten, die nur Text enthalten.

Bei den Ergebnissen mit der neuen Bildähnlichkeitssuche kann festgestellt werden, dass die Erkennungsleistung sehr viel besser ist. Es werden auch Illustrationen gefunden, die unterschiedliche Druckstärke aufweisen und enger zugeschnitten sind.Darüber hinaus ist anzumerken, dass die Bildähnlichkeitssuche mit der neuen Lösung erheblich schneller ist als die alte Lösung. Diese Beobachtung konnte durch eine Vielzahl an Experimenten und weiteren Ähnlichkeitssuchen bestätigt werden.

Gewonnene Erkenntnisse bei der neuen Bildähnlichkeitssuche

Durch das komplette Redesign der Bildähnlichkeitssuche ergaben sich einige Herausforderungen, die gemeistert werden mussten. Zunächst wurde untersucht, ob die extrahierten Bildsegmente der alten Bildähnlichkeitssuche als Grundlage für das Training eines Klassifikators zur Erkennung von Bildsegmenten verwendet werden können. Diese Idee wurde verworfen, nachdem eine genaue Fehleranalyse der extrahierten Bildsegmente vorgenommen wurde. Dabei wurde festgestellt, dass sehr viel Text oder gar ganze Scans als Bildsegmente vorhanden sind. Deshalb wurde das Training des Klassifikators auf Basis von manuell annotierten Bildsegmenten vorgenommen, um diverse Fehlerquellen im Vorhinein ausschließen zu können.

Beim Training eines Klassifikators war zunächst nicht klar, wie granular die notwendigen Bildsegmente annotiert werden sollten, also ob beispielsweise Karten noch einmal in kleinere Segmente unterteilt werden sollten. Des Weiteren war am Anfang nicht bekannt, wie genau sich Annotationsfehler auf das spätere Erkennen von Bildsegmenten auswirken. So führte beispielsweise das Annotieren von Tabellen als Bildsegment dazu, dass Seitenzahlen in Büchern oder andere Textstellen als Segmente erkannt wurden. Das hat den nachteiligen Effekt, dass so u. a. der Index an Vektoren unnötig groß wird, weil viele Segmente im Index gespeichert sind, die eigentlich keine sinnvollen Bildsegmente sind.

Eine weitere Herausforderung zeigte sich beim Training des Klassifikators. Da künstliche neuronale Netze zum Einsatz kommen, ist das Training auf Standardhardware sehr zeitintensiv, sodass auf Spezialhardware (GPUs) zurückgegriffen wurde. Es konnte außerdem festgestellt werden, dass die Menge an annotierten Daten nicht ausreicht, um Transformer-basierte Modelle für die Erkennung von Bildsegmenten zu verwenden. Deshalb wurde ein CNN-basiertes Modell für die Erkennung von Bildsegmenten verwendet, das nichtsdestotrotz eine hervorragende Erkennungsleistung zeigt und darüber hinaus schneller trainiert werden kann.

Auch beim Anwenden des Klassifikators auf unseren Bestand, um Bildsegmente zu extrahieren, kam Spezialhardware zum Einsatz. Das sog. Taggen des Bildbestandes kann je nach Größe des Bestands mehrere Wochen bzw. Monate dauern, selbst wenn Spezialhardware zum Einsatz kommt. Der Begriff »Taggen« bezeichnet hier das automatische Annotieren, also den Prozess, der vorher von Menschen gemacht wurde. Das effiziente Ausnutzen dieser Spezialhardware kann ebenfalls als Herausforderung angesehen werden, da hierbei mehrere Versuche und Benchmarks unternommen wurden, um die Speicherkapazität der GPUs komplett auszunutzen.

Fazit

Der dargestellte Weg der Bayerischen Staatsbibliothek von der »alten«, bereits 2013 implementierten Bildähnlichkeitssuche zum aktuellen Neuansatz, der unter anderem auf künstlichen neuronalen Netzen basiert, macht deutlich, wie wichtig es auf dem Feld innovativer digitaler Kulturtechnologien ist, sich frühzeitig als Pionier zu positionieren. Nur so lässt sich sukzessive ein anwendungsorientierter Erfahrungsschatz erwerben, der ein »Weitermachen« entlang innovativer Neuentwicklungen überhaupt erst möglich macht. Mit lediglich projektförmig aufgesetzten »Konzeptstudien« und »Prototypen« wird man hier (leider) nicht weit kommen. Und es zeigt sich wieder einmal, dass fortgeschrittene Technologien, wie die hier beschriebene Applikation, schon signifikante Massen an digitalen Inhalten brauchen, um überhaupt performant operieren und wissenschaftsrelevante Resultate erzeugen zu können. Insofern bestätigt sich einmal mehr das strategische Mantra der Bayerischen Staatsbibliothek: Content in Context.

Anmerkungen

1 https://bildsuche.digitale-sammlungen.de
2 Siehe die Darstellung zum Projekt: Markus Brantl, Klaus Ceynowa, Thomas Meiers, et al., Visuelle Suche in historischen Werken, in: Datenbank Spektrum 17 (2017) S. 53–60.
3 Martin Hermann, Ralf Eichinger und Barbara Niegisch: IIIF Reloaded: Die neuen Digitalen Sammlungen der Bayerischen Staatsbibliothek, in: ABI-Technik 41,4 (2021) S. 244–254.
4 Projektwebseite unter: https://projectmirador.org/
5 Siehe dazu die GitHub-Präsenz des MDZ: https://github.com/dbmdz/solr-ocrhighlighting
6 Projektwebseite unter: https://prodi.gy/
7 https://www.image-net.org/index.php
8 Kaiming He, Xiangyu Zhang, Shaoqing, et al., Deep Residual Learning for Image Recognition, in: IEEE Conference on

Computer Vision and Pattern Recognition (CVPR) (2016), S. 770–778.

9 Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., Attention Is All You Need, in: Advances in Neural Information Processing Systems 30 (NIPS 2017).

10 Google AI Blog: Recent Advances in Google Translate [online] [Zugriff am: 25.08.2022]. Verfügbar unter: https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

11 Ein sehr bekanntes Sprachmodell ist BERT: Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (2019) S.4171–4186.

12 Sprachmodelle des Münchener Digitalisierungszentrums stehen Open Source zur Verfügung unter: https://github.com/dbmdz/berts

13 Ze Liu, Yutong Lin, Yue Cao, et al., Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021).

14 Hugo Touvron, Matthieu Cord, Matthijs Douze, et al., Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning (2021) S. 10347–10357.

15 Kunihiko Fukushima, Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position, in: Biological Cybernetics (1980) S. 193–202.

16 John Denker, W. Gardner, Hans Graf, et al., Neural Network Recognizer for Hand-Written Zip Code Digits, in: Advances in Neural Information Processing Systems 1 (NIPS 1988).

17 Yann Lecun, Léon Bottou, Yoshua Bengio, et al., Gradient-Based Learning Applied to Document Recognition, in: Proceedings of the IEEE (1998) S. 2278–2324.

18 Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems 25 (NIPS 2012).

19 Projektwebseite unter: https://github.com/facebookresearch/detectron2

20 Überblick über Active Learning Strategien: Burr Settles, Active Learning Literature Survey [online] [Zugriff am: 25.08.2022]. Verfügbar unter: https://burrsettles.com/pub/settles.activelearning.pdf

21 Projektwebseite unter: https://solr.apache.org/

22 Versionschangelog unter: https://solr.apache.org/docs/9_0_0/changes/Changes.html

Verfasser

Dr. Markus Brantl, Abteilungsleiter Digitale Bibliothek und Bavarica, Bayerische Staatsbibliothek, Ludwigstraße 16, 80539 München, markus.brantl@bsb-muenchen.de, https://orcid.org/0000-0003-0967-8064

Foto: Bayerische Staatsbibliothek / H.-R. Schulz

Stefan Schweter, Softwareentwickler und Experte für KI-Systeme am Münchener Digitalisierungszentrum, Bayerische Staatsbibliothek, Ludwigstraße 16, 80539 München, stefan.schweter@bsb-muenchen.de, https://orcid.org/0000-0002-7190-2090

Foto: Bayerische Staatsbibliothek / D. Khromeev