Dominik Bönisch • Francis Hunger

»THE CURATOR'S MACHINE«

KORRELATIONEN IN DEN NETZWERKEN KÜNSTLICHER ›INTELLIGENZ‹ IM VERGLEICH ZU DATENBANKANWENDUNGEN

Anhand des Forschungsprojektes »Training the Archive« diskutiert der folgende Text den korrelativen Charakter von Informationen, die mit Hilfe Künstlicher ›Intelligenz‹ (KI) in gewichteten Netzwerken verarbeitet werden.[1] Es geht zentral um die Frage, wie Sammlungs- und Netzwerklogiken in der auf KI basierenden explorativen Software »The Curator's Machine« anders ineinandergreifen, als dies bei reinen Datenbankanwendungen der Fall ist.

Im Fokus von »Training the Archive« stehen die Möglichkeiten und Risiken von KI in Bezug auf die kuratorische Praxis. Bis Ende 2023 soll erprobt werden, inwieweit KI für die Erkundung musealer Sammlungsdaten eingesetzt werden kann. Daraus leitet sich die Forschungsfrage ab, ob es möglich ist, einer KI-Software den kuratorischen Rechercheprozess anzutrainieren, so dass mit Hilfe von maschinellem Lernen[2] Zusammenhänge zwischen Kunstwerken offenbar werden, die für den Menschen nicht unmittelbar ersichtlich sind. Dieser Frage geht das Projekt »Training the Archive« seit Anfang 2020 nach.

1 Im vorliegenden Text werden die Begriffe der ›Künstlichen Intelligenz‹ und des ›Neuronalen Netzwerks‹ mit Vorsicht eingesetzt. Um den Diskurs zu deanthropomorphisieren, wird auf ›Künstliches Neuronales Netzwerk‹ (KNN) zugunsten von ›gewichtetem Netzwerk‹ verzichtet, da die Knoten dieser Netzwerke zwar historisch durchaus biologisch inspiriert waren, in der heutigen technischen Funktion jedoch mit neuronalen Strukturen wenig gemein haben. Daneben verzichtet der Text auf Verben wie ›lernen‹ oder ›erkennen‹, welche menschliche Geistesleistungen voraussetzen würden. Siehe Francis Hunger: Unhype Artificial Intelligence! A proposal to replace the deceiving terminology, in: Training the Archive – Working Paper 6, Aachen/Dortmund, 04/2023, https://zenodo.org/record/7524493.

2 ›Maschinelles Lernen‹ ist ein feststehender Begriff, der allerdings eher eine statistisch-mathematische Optimierung beschreibt als tatsächliches Lernen im menschlichen Sinne. Maschinelles Lernen umfasst die Entwicklung eines Modells mittels Algorithmen, welche auf eine große Menge an Trainingsdaten zurückgreifen. Die in dem Modell gespeicherten statistischen Verteilungen können für die Erstellung von Vorhersagen oder Empfehlungen genutzt werden.

DOI https://doi.org/10.15499/kds-006-011 | CC BY-SA 4.0

Das Ludwig Forum für Internationale Kunst in Aachen ist ein zeitgenössisches Kunstmuseum mit Fokus auf der Sammlung Ludwig, die am Standort über 4.000 Werke internationaler Kunst ab den 1960er-Jahren umfasst. »Training the Archive« ist ein Verbundprojekt mit dem HMKV Hartware MedienKunstVerein in Dortmund – einem Kunstverein mit langjähriger Expertise im Bereich der Medienkunst. Für die Entwicklung von passenden Softwareprototypen arbeitet »Training the Archive« mit dem Visual Computing Institute der RWTH Aachen University zusammen. Neben der Erarbeitung eines Softwareprototyps etablierte »Training the Archive« eine Ebene der wissenschaftlichen Reflexion mittels sogenannter Working Papers, welche die Erkenntnisse zur Diskussion stellen. Zu den Projektergebnissen zählen weiter die Durchführung einer Konferenz zum Thema »Kunst & Algorithmen« am 17. und 18. November 2022 im Ludwig Forum Aachen, zehn Videointerviews mit einschlägigen Expert*innen auf YouTube, eine laborhafte Präsentation im Museum sowie eine Abschlusspublikation.[3]

Im ersten Teil dieses Beitrags wird die Entwicklung der Software, die »Curator's Machine«, dargestellt, im zweiten Teil werden medientheoretische Überlegungen anhand der besprochenen Prototypen zum Verhältnis von Netzwerken, strukturierten Informationen und Sammlungen entwickelt. Zunächst sind aber einige grundsätzliche Erläuterungen zu den verwendeten Begrifflichkeiten, zur oben aufgeworfenen Problemstellung und ihrem Kontext zu geben.

In relationalen Datenbanken liegen Daten strukturiert vor und diese Ordnung fungiert als Träger semantischer Informationen. Im Unterschied dazu sehen Convolutional Neural Networks,[4] im Folgenden als ›gewichtete Netzwerke‹ bezeichnet, von der Struktur ab und prozessieren Bilddaten auf Pixelebene. Dabei werden innerhalb der Netzwerke bestimmte Gewichtungen verstärkt und andere abgeschwächt. Diese Gewichte innerhalb eines solchen Netzwerks bestimmen, mit welcher Wahrscheinlichkeit ein Aktivierungssignal von einem Knoten zum nächsten weitergeleitet wird. Schließlich kann für jedes trainierte Bild ein statistisch-mathematischer Vektor erzeugt werden, der Korrelationen zwischen Bildern beschreibt und Gruppierungen

3 Das Projekt wurde 2020 bis 2023 im Programm »Kultur Digital« der Kulturstiftung des Bundes von der Beauftragten der Bundesregierung für Kultur und Medien gefördert.

4 Convolutional Neural Networks sind algorithmische Assemblagen mit sehr vielen Parametern, die untereinander verknüpft sind und somit ein Netzwerk ergeben. Dieses Netzwerk ist in zahlreichen Schichten aufgebaut. ›Convolutional‹ bezieht sich auf ein spezifisches algorithmisches Verfahren, wie die einzelnen Netzwerkknoten aktiviert werden.

oder Klassifikationen ermöglicht. Die Gesamtheit der in den gewichteten Netzwerken prozessierten Vektoren wird in einen latenten Raum projiziert. Dieser kann mittels Algorithmen rechnerisch durchschritten werden, verweigert sich allerdings aufgrund seiner Mehrdimensionalität der unmittelbaren Einsicht, wie dies in Tabellen und Datenbanken der Fall ist.[5] So entsteht in Korrelationsnetzwerken im Vergleich zu strukturierten Daten in relationalen Datenbanken eine ›unscharfe‹ Datenverarbeitung.

Aktuelle technologische Entwicklungen (zum Beispiel das CLIP-Modell) erlauben es, Zugehörigkeit als Korrelationen sowohl auf der Pixelebene der Bilder als auch durch multimodale Vektoren der Metadaten zu berechnen.[6] Referenz tritt in einer spezifischen Form auf, und zwar nicht unmittelbar durch diagrammatische Verteilungen der Information (wie in Tabellen und Datenbanken), sondern mittelbar als Korrelation mathematischer Vektoren in einem Wahrscheinlichkeitsraum. Diese Zweiteilung der Informationsverarbeitung in 1.) ›strukturiert‹ und 2.) ›korrelativ‹ wird in aktuellen KI-Anwendungen zusätzlich verkompliziert. Denn die latenten Räume in den korrelativen Netzwerken der Computer Vision werden in den User-Interfaces erneut mit strukturierten Sammlungsdaten, den Metadaten, überlagert.[7]

5 Ein Vektor beschreibt über Zahlenwerte Position und Orientierung im Raum. Vektoren erlauben daher das Rechnen in räumlichen Verhältnissen. Im Schulunterricht wird mit zwei- oder dreidimensionalen Vektoren gerechnet. Gewichtete Netzwerke hingegen bauen multidimensionale Vektoren, die als Features bezeichnet werden, auf. Sie erlauben es, die Nähe oder Entfernung dieser mehrdimensionalen Vektoren zu berechnen. Daraus ergeben sich Ähnlichkeiten, denn Vektoren, die einander näher liegen, sind statistisch gesehen ähnlicher, als weiter entfernte. Anhand der Ähnlichkeiten ergeben sich Cluster von zugehörigen Vektoren. Spezifikum dieses korrelativen, latenten Raums ist es, dass einzelne Datenpunkte beziehungsweise Vektoren immer nur indirekt aus anderen beobachtbaren Variablen abgeleitet werden können und zwar über ein mathematisches Modell. Diese Datenpunkte sind im Prinzip Thesen, daher wird der mit ihnen aufgebaute multidimensionale Raum als ›latent‹ bezeichnet.

6 CLIP (Contrastive Language – Image Pre-Training) ist ein vortrainiertes gewichtetes Netzwerk, welches anhand statistischer Funktionen, d.h. anhand von Vektor-Korrelationen, Text-Bild-Paare berechnet. Somit ist es möglich, Bildern (hier gedacht als Anordnungen von Pixeln) bestimmte Text-Phrasen (hier konzipiert als statistische Verteilungen von Worten) zuzuordnen. Eine gut verständliche Erklärung der technischen Funktionsweise von CLIP leistet Hannes Bajohr: Dumme Bedeutung. Künstliche Intelligenz und artifizielle Semantik, in: Merkur 76/882 (2022), S. 69–79, hier S. 76.

7 Auch wenn einzelne Konzepte im vorliegenden Text in den Fußnoten erklärt werden, so setzt der Text doch Grundkenntnisse der Funktionsweisen gewichteter

Diese grundsätzlichen Erwägungen flossen auch in die Konzeption und Umsetzung des Projekts »Training the Archive« ein. Bereits vorab zeigte sich, dass die Arbeit mit ephemeren, zeitbasierten, performanceorientierten oder dreidimensionalen Kunstwerken schwierig werden würde, daher konzentrierte sich »Training the Archive« auf zweidimensional abbildbare Kunstwerke beziehungsweise auf deren Entsprechung als zweidimensionale Digitalisate. Die durch die Digitalisierung von Kunstobjekten entstandenen ›operationalen‹ Ab-Bilder[8] unterliegen zahlreichen Beschränkungen hinsichtlich der Erfassung des künstlerischen Inhalts.[9] Als Eingabedaten spiegeln sie weder Materialitäten noch konzeptionelle Strategien wider, welche, wenn überhaupt, durch die Beigabe von Metadaten erfasst werden können. In den Prototypen von »Training the Archive« zeigt sich somit eine Entwicklung: Während die ersten Prototypen auf mit ImageNet vortrainierten Netzwerken basierten, griff der dritte – und für das Projekt finale – Prototyp auf das vortrainierte, multimodale CLIP-Modell zurück. Durch ein User-Interface, welches auch Metadaten einbezieht, wurden zudem erneut strukturierte Daten eingewoben.

Obwohl die verwendeten Machine-Learning-Modelle in der Lage sind, visuelle Anordnungen von Kunstwerken zu erzeugen, sollte in »Training the Archive« darüber hinaus ein Prozess der Mensch-Maschine-Interaktion etabliert werden. Somit würde auch das historische, stilistische und objektbasierte Kontextwissen von Expert*innen, wie zum Beispiel Kurator*innen, einbezogen werden. Ziel war also nicht das Hinwegautomatisieren des Menschen, sondern eine Betonung des ›Human-in-the-Loop‹ und eine ›Kol-

Netzwerke voraus. Eine grundsätzliche Einführung in Machine Learning mit visuellen Datensätzen leistet Jenna Burrell: How the machine ›thinks‹ – Understanding Opacity in Machine Learning Algorithms, in: Big Data & Society 3/1, 5.1.2016. Einige der Voraussetzungen werden in den Working Papers 1 und 2 diskutiert. Siehe Dominik Bönisch: The Curator's Machine. Clustering von musealen Sammlungsdaten durch Annotieren verdeckter Beziehungsmuster zwischen Kunstwerken, in: Training the Archive – Working Paper 1, Aachen/Dortmund, 10.5.2021, https://zenodo.org/record/4604880; Francis Hunger: ›Why so many windows?‹ – Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst, in: Training the Archive – Working Paper 2, Aachen/Dortmund, 1.6.2021, https://zenodo.org/record/4742621.

8 Harun Farocki: Der Krieg findet immer einen Ausweg, in: Essay – Cinema 50, hg. von Natalie Böhler, Marburg, 2005, S. 21–33.

9 Vgl. Beryl Graham und Sarah Cook: Rethinking Curating – Art after New Media, Cambridge, MA 2010; Oliver Grau, Janina Hoth und Eveline Wandl-Vogt: Digital art through the looking glass – new strategies for archiving, collecting and preserving in digital humanities, Krems a.d. Donau 2019.

laboration‹ zwischen Mensch und Maschine.[10] Dabei wird die »Curator's Machine« als eine iterative Sammlung von Prototypen – aber auch theoretischen Konzepten – verstanden, die sich im Laufe des Forschungsprojektes weiterentwickelt hat. Im Folgenden gilt es, diese Entwicklung am Beispiel dreier Softwareprototypen genauer zu beschreiben.

Prototyp I – Annotieren verdeckter Beziehungsmuster zwischen Kunstwerken

Für den ersten Prototyp der »Curator's Machine« sollten digitalisierte Abbildungen von Kunstwerken mittels eines gewichteten Netzes prozessiert werden, um den Datensatz automatisch in verschiedene Gruppen sortieren zu lassen (Abb. 1). Dabei erfolgte das Clustering[11] anhand von visuellen sowie technischen Merkmalen, den Features beziehungsweise Vektoren, die den Netzen bereits eintrainiert sind. Das können zum Beispiel Farbwerte, Texturen oder Formen und Teilobjekte im Bild sein. Daraus ergibt sich eine maschinell-statistische ›Interpretation‹ der digitalen Abbildungen von Kunstwerken, welche für die Gruppierung von Bildern mit ähnlichen Merkmalen innerhalb des mehrdimensionalen Raumes gewichteter Netze genutzt werden können. Schließlich sollte es möglich werden, den Datensatz durch zweidimensionale Projektionen wie Raster- oder Streudiagramme zu visualisieren, wobei sich ähnliche Bilder zu Clustern zusammenfinden, als sogenannte ›Nearest Neighbors‹. Die Zueinanderordnung von Bildern anhand ähnlicher Features wirft einen neuen Blick auf die Sammlung, da die mittels Metadaten etablierte Ordnung ignoriert beziehungsweise neu verhandelt wird.

Mittels Clustering werden Muster und Zusammenhänge an der Bildoberfläche erkannt. Dieses Oberflächen-Clustering ist inzwischen in einer Reihe von Projekten erforscht worden (zum Beispiel »Imgs.ai«, »iArt« und »Google Images«), hält jedoch weiterhin Überraschungen bereit. »Training the Archive« untersuchte, ob mittels Machine Learning auch ›Verborgenes‹ unterhalb der Bildoberfläche prozessiert werden kann und ob zum Bei-

10 Dominik Bönisch: Suggestions for a Curator's Machine: A Collaborative Approach to the Use of Artificial Intelligence in Art Museums, in: Art, Museums and Digital Cultures – Rethinking Change, Lisabon: Instituto de História da Arte, Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa, 2021, S. 136–148.

11 Ausführlich zum Clustering als diagrammatisches Werkzeug, siehe Francis Hunger: Punktwolken. Scatterplots Und Tabellen Als User-Interfaces Künstlicher ›Intelligenz‹, in: Training the Archive – Working Paper 5, Aachen/Dortmund (23.1.2023), https://zenodo.org/record/7554463.

Abb. 1: Clustering als Scatterplot anhand von Bild-Ähnlichkeiten (im Sinne von Pixel-Ähnlichkeiten)

spiel verdeckte Beziehungsmuster zwischen Kunstwerken sichtbar gemacht oder persönliche Intuition und subjektiver Geschmack von verschiedenen Kurator*innen modelliert werden können.

Im ersten Experiment hat unser Projekt demnach geprüft, ob die durch die »Curator's Machine« gebildeten Cluster mittels eines ›Trainings‹ über menschengemachte, zusätzliche Annotationen – und zwar Information über die Zusammengehörigkeit von Kunstwerken – verändert werden können. Mit diesen Zusammengehörigkeitsinformationen wurde ein Netzwerk trainiert, so dass sich darin die Gewichtungen entsprechend veränderten. In diesem Experiment wurden die Zuordnungen von Bildern in Clustern mithilfe des ›Triplet-Loss-Verfahrens‹ beeinflusst. Durch zusätzliche Gewichtungen, die

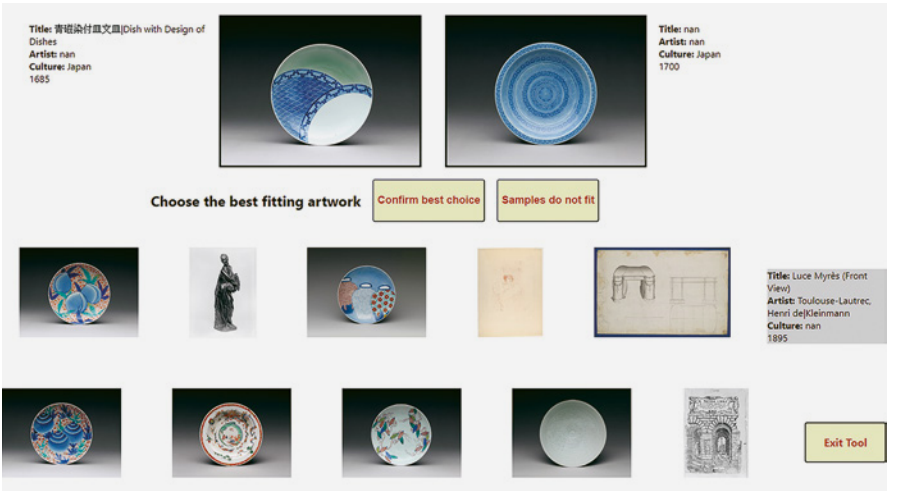

Abb. 2: Workflow für die Auswahl von passenden Kunstwerken. Mit Hilfe der Buttons ›Confirm best Choice‹ und ›Samples do not fit‹ kann die Auswahl verändert werden, was als zusätzliche Gewichtungen in das Netzwerk eingeschrieben wird.

als Triplets (negative Gewichtung – Ankerwert – positive Gewichtung) organisiert sind, können Nutzer*innen der Software zeigen, was aus ihrer Sicht zusammengehörig ist oder nicht (Abb. 2).[12] Dieses Prinzip wurde auch in spätere Prototypen übernommen und stellte somit einen wichtigen Baustein für die finale »Curator's Machine« dar.[13]

Einschränkend gilt es zu bedenken, welches Wissen hierbei Beachtung findet, von wem dieses stammt und zu welchem Anteil es einfließt, um die Reproduktion von Bias, wenn möglich, zu vermeiden. So geschieht die Feature-Extraktion mittels gewichteter Netze (wie »InceptionV3«, »BiT/m-r152×4« und »VGG19«), welche mit der Bilddatensammlung ImageNet[14] vortrainiert sind. Diese unterliegen gleich mehreren Bias:

12 Die Verwendung von Triplets ist auf eine Idee von Benoît Seguin zurückzuführen: Benoît Laurent Auguste Seguin: Making large art historical photo archives searchable, Dissertation, EPFL, Lausanne 2018, S. 65.

13 Vgl. D. Bönisch (wie Anm. 7).

14 Fei-Fei Li u.a.: ImageNet: A Large-Scale Hierarchical Image Database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Juni 2009, S. 248–255; Geoffrey E. Hinton, Alex Krizhevsky und Ilya Sutskever: ImageNet Classification with Deep Convolutional Neural Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems – Volume 1, Curran Associates Inc., Redhook, NY 2012, S. 1097–1105.

1. Textur-Bias: Bei dem aufgrund der eingesetzten mathematischen Verfahren die gewichteten Netze Ähnlichkeiten anhand von Texturen besser detektieren, als anhand von Umrissen.[15]
2 Ontologischer Bias: Dieser wird durch die ImageNet zugrundeliegenden Ontologien von WordNet eingeführt,[16] welches seit einiger Zeit nicht mehr aktiv weiterentwickelt wird und beispielsweise problematische Kategorien wie ›Schönheit‹ oder ›sexuelle Orientierung‹ in den Bildannotationen zur Verfügung stellte, so als ob es für Schönheit ein objektives Maß gäbe, oder als ob die sexuelle Orientierung aus dem Bild eines Menschen abgelesen werden könne.[17]
3. Daten-Bias: Die unzureichende demographische Präzision der ImageNet-Datenbank, in der zum Beispiel im Vergleich zu den tatsächlichen prozentualen Verteilungen der Populationen übermäßig viele weiße Männer enthalten sind.[18]
4. Bias aufgrund von A-Historizität: Das Einebnen jeglicher Zeitlichkeit, so dass beispielsweise ein Stichwort wie ›Faltenwurf‹ unterschiedslos Bilder aus dem 16. Jahrhundert und zeitgenössische Abbildungen von Bürovorhängen aus dem Ikea-Katalog betrifft.[19]

Wie stark diese Bias nicht nur den Output gewichteter Netze beeinflussen, sondern auch bei der durch »Training the Archive« verwendeten Feature-Extraktion wirksam wurden, ist derzeit eine offene Forschungsfrage. Bisher ist davon auszugehen, dass sie mit den extrahierten Features übertragen werden. Wir sehen hier bereits, dass nicht allein die Strukturen gewichteter Netze wichtig sind, sondern auch die durch sie prozessierten Daten. Die Herstellung von Daten und ihren Ordnungen, in denen sie an die »Curator's Machine« übergeben werden, spielen daher eine nicht zu unterschätzende Rolle.

15 Robert Geirhos u. a.: ImageNet-trained CNNs are biased towards texture – Increasing shape bias improves accuracy and robustness, in: arXiv:1811.12231, 14. 1. 2019, http://arxiv.org/abs/1811.12231.
16 George A. Miller u. a.: Introduction to WordNet: An On-line Lexical Database, in: International Journal of Lexicography, 3/4, 1990, S. 235–244.
17 Fei-Fei Li u. a.: Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 27. 1. 2020, S. 547–558, hier S. 2.
18 Joy Buolamwini und Timnit Gebru: Gender Shades – Intersectional Accuracy Disparities in Commercial Gender Classification, in: Proceedings of Machine Learning Research – Conference on Fairness, Accountability, and Transparency, 81, 2018, S. 1–15, hier S. 12.
19 F. Hunger (wie Anm. 7); Gabriel Pereira und Bruno Moreschi: Artificial intelligence and institutional critique – Unexpected ways of seeing with computer vision, in: AI & Society, 14. 9. 2020.

Zusammenfassend lässt sich sagen: Der erste Prototyp der »Curator's Machine« ermöglichte Kurator*innen das zusätzliche Annotieren verdeckter Beziehungsmuster zwischen Abbildungen von Kunstwerken.[20] Da sich jedoch das Vergeben der Annotationen als arbeitsintensiv erwies – es mussten für ein Datenkorpus bis zu 3.000 solcher zusätzlichen Annotationen trainiert werden –, stellte sich die Frage nach der Praktikabilität im kuratorischen Alltag. Für einen zweiten Prototypen galt es daher, einen anderen Ansatz zu wählen.

Prototyp II – Empfehlungssystem, basierend auf Trajektorien von Bildauswahl und Nearest Neighbors

Der zweite Prototyp (Abb. 2) war ein topologisches Empfehlungssystem, das auf Basis des K-Nearest-Neighbors-Prinzips zu einer Auswahl von Objekten aus einer digitalisierten Sammlung passende Kunstwerke vorschlägt. Das Verfahren bezieht sich historisch auf das sogenannte Brückenproblem, welches 1736 durch Leonhard Euler mathematisch untersucht wurde. Euler fand einen Algorithmus für den kürzesten Weg zwischen zwei Punkten in der Stadt Königsberg, während in seiner Überlegung eine Reihe von Brücken überquert werden mussten, und begründete damit ein Forschungsfeld, welches später als Graphentheorie bekannt wurde.[21] Nach diesem Prinzip wurde für den zweiten Prototyp der Pfad zur Exploration des latenten Raumes[22] innerhalb eines eigens aufgebauten Autoencoder-Netzes berechnet.[23] Das Empfehlungssystem wurde nun darauf trainiert, die Beziehung zwischen

20 Der Prototyp ist online dokumentiert unter: https://github.com/DominikBoenisch/Training-the-Archive.

21 Siehe L. Euler in: Dénes König: Theorie der endlichen und unendlichen Graphen, hg. von Horst Sachs und H. Sachs, Teubner-Archiv zur Mathematik 6, 1936, Reprint, Leipzig 1986, S. 290–301.

22 Zur Begriffsklärung ›Latenter Raum‹ beziehungsweise ›hochdimensionaler Raum‹, siehe Anm. 5.

23 Der Autoencoder verarbeitet Bilddaten durch einen ›Flaschenhals‹, wodurch das gewichtete Netzwerk gezwungen wird, High-Level-Features wie Form, Farbe oder Stil von Bildern zu vergleichen. Anhand der Features, d.h. anhand von Pixelähnlichkeiten, werden die Sammlungsbilder angeordnet. Ergänzt werden diese ›Embeddings‹ durch Metadaten zu den Kunstwerken, so dass sowohl die Pixelwerte des Bildes als auch bestimmte Informationen (zum Beispiel Titel, Werkgröße) Ähnlichkeitsbeziehungen ergeben. Wie Embeddings in den latenten Raum gewichteter Netzwerke mittels ›unsupervised learning‹ eingeschrieben werden, erklärt anschaulich H. Bajohr (wie Anm. 6), hier S. 73.

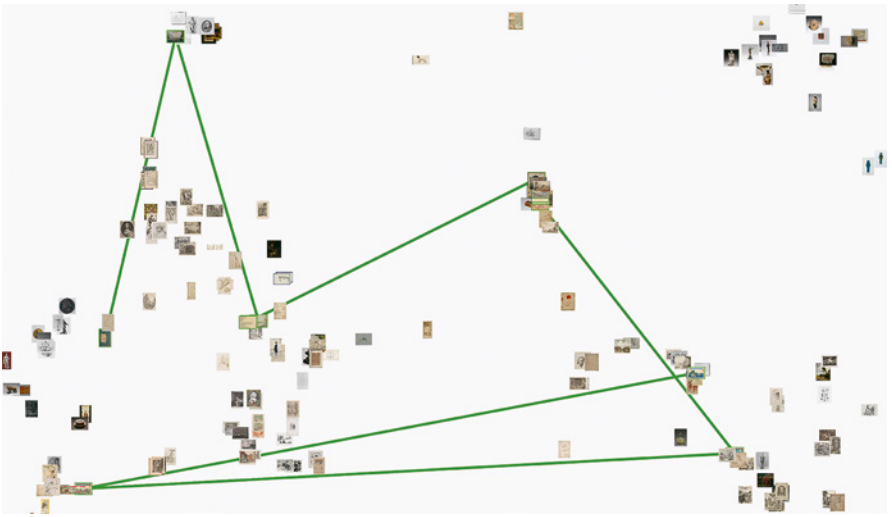

Abb. 3: Trajektorie als Scatterplot des Pfades durch den hochdimensionalen Raum von einem Kunstwerk zu einem anderen

verschiedenen Kunstwerken zu berücksichtigen, indem Expert*innen eine Sequenz von Kunstwerken auswählten, die in einer Ausstellung zusammengehören könnten. Diese annotierte Sequenz ergab einen Pfad durch den Einbettungsraum, die das Empfehlungssystem statistisch nachbildete, indem es weitere Vorschläge kalkulierte, um den ›Pfad‹ fortzusetzen und indem es statistische Nearest Neighbors zur Auswahl stellte (Abb. 3).

Der wissensbildende Aspekt für die Nutzer*innen eines solchen Systems liegt darin, bisher weniger beachtete Werke aus der Sammlung zu entdecken und somit eigene kuratorische Vorannahmen und Grundhaltungen zu überprüfen und potenziell zu erweitern. Dieser Prototyp zählt somit zu den explorativen Systemen, die ein ›Neu-Kennenlernen‹ der Sammlung ermöglichen.

Einschränkend ist zu erwähnen, dass das Verfahren von der Annahme ausging, dass für die Annotation der Pfade eine Ausstellung streng sequenziell aus Exponaten zusammengestellt wird, die einander ähnlich sind oder auf Metadatenbasis zusammenhängen. In Wirklichkeit ist diese Kontinuität nicht immer der Fall, insbesondere wenn: a) thematische Brüche in eine Ausstellung eingebracht werden, b) Themenwechsel die Reihenfolge von Werken ändern oder c) dialektische Beziehungen zwischen Arbeiten aufgerufen werden sollen. Der Prototyp II schlug ähnliche Bildobjekte vor, jedoch basiert die kuratorische Auswahl nicht allein auf Ähnlichkeiten: War beispielsweise

ein bestimmter Sammlungsgegenstand ausgewählt, war nicht zwangsläufig ein weiteres Bild der gleichen Objektklasse gesucht.

Aus diesen Einschränkungen und Erkenntnissen, die sich im Laufe der Arbeit am zweiten Prototyp ergaben, folgte die Notwendigkeit, die Vorgehensweise noch einmal neu aufzustellen. Bevor hierfür neue konzeptuelle Überlegungen angestellt wurden, galt es, mehr empirisches Material zur kuratorischen Arbeit zu sammeln.

Empirisches Material – Kurator*inneninterviews

Um die Komplexitäten kuratorischer Tätigkeit besser zu erfassen, wurde eine Reihe strukturierter Interviews mit Kurator*innen durchgeführt. Zu den befragten Expert*innen zählten: Sabine Maria Schmidt (Kunstsammlung Chemnitz), Severin Dünser (Belvedere 21, Wien), Joasia Krysa (Liverpool Biennial), Marie Lechner (unabhängige Kuratorin, Paris), Daniel Muzyczuk (Muzeum Sztuki Łódź, Polen), Janice Mitchell (Museum Ludwig, Köln), Tina Sauerländer (unabhängige Kuratorin, Berlin), Raffael Dörig (Kunsthaus Langenthal, Schweiz), Kasia Redzisz (KANAL – Centre Pompidou, Brüssel), Xiaoyu Weng (The Guggenheim Museums, New York) und Anna Fricke (Museum Folkwang, Essen). Die codierten Transkripte erlaubten, die Prinzipien der kuratorischen Praxis besser zu verstehen und die Bestandteile dieses komplexen Prozesses aufzudecken, die als formalisierbare wissensbildende Verfahren in Machine-Learning-Modelle einfließen können.[24] Insbesondere Aussagen zur Werkauswahl, zum Zusammenstellen von Künstler*innenpositionen und zu Ausstellungsideen und -konzepten gingen in die Überlegungen zum weiteren Verfahren ein.

Als erstes Ergebnis zeigte sich, dass Kurator*innen verstärkt nach den Kontexten von Kunstwerken fragen und Ausstellungen als inhaltlich verknüpfte Zusammenstellungen erarbeiten. Eine Kuratorin beschrieb die Recherche als eine Reihe von Hyperlinks: »I mean, going from a link to another to another to another, and try to associate things that I found, say in very different spaces.« Ein anderer Kurator hob Gruppierungen und Konstellationen hervor: »Art historical research exhibitions usually group types. This is basically usually uncovering an unknown material or not frequently looked at material. Or trying to put this material into a constellation that reactivates

24 Die Codierung erfolgte in einer Mischung aus deduktiver und induktiver qualitativer Analyse nach Uwe Flick, Ernst von Kardorff und Ines Steinke (Hg.): Qualitative Forschung – ein Handbuch, Reinbek bei Hamburg 2022.

it.« Eine weitere Kuratorin beschrieb ihre Vorgehensweise als visuelles Clustern: »And then we interrogate those clusters, like making sure that it makes sense all together. There is a visual aspect to it; formal aspects to it.«[25] In all diesen Zitaten spiegeln sich Praktiken, welche dem ordnenden Herstellen von Zusammenhängen dienen, sodass das Kuratieren als Kontextualisierungs- und Ordnungspraxis aufgefasst werden kann.

Es zeigten sich in den Interviews auch eine ganze Reihe kuratorischer Tätigkeiten, die nicht Teil der »Curator's Machine« wurden: Aufgaben wie beispielsweise die konkrete Installation im Ausstellungsraum, Logistik und Management, Finanzierung, Ausstellungsvermittlung und innerinstitutionelle Aspekte konnten in dem Softwareprototypen nicht berücksichtigt werden. Diese sind teilweise durch andere digitale Anwendungen und Verfahren bereits unterstützt oder zu komplex, als dass sie in der »Curator's Machine« abgedeckt werden könnten. Immer deutlicher stellte sich heraus, dass die »Curator's Machine« eine interaktive Suchmaschine werden sollte, die in der Lage ist, anhand von Visualitäten zu suchen, und die mit semantischen Daten angereichert wird, um außerdem auch Text-Bild-Assoziationen zu berücksichtigen. Es ergab sich die Frage, welche technischen Modelle und welche prozessualen Verfahren es ermöglichen könnten, den kuratorischen Kontextualisierungsverfahren noch besser gerecht zu werden.

Prototyp III – Eine interaktive Suchmaschine

Um Kunstwerke in Sammlungen zugänglicher zu machen, erforschte »Training the Archive« Wege, die rein visuelle Suche der bisherigen Prototypen multimodal zu erweitern. Als ›Multimodal Vision-Language Models‹ werden gewichtete (Transformer-)Netze bezeichnet, welche die Datenmengen in mehreren Modi ansprechen können, in diesem Fall über den Modus der Computervision und den der Texterkennung. Dies erfolgt durch Bild-Text-Korrelationen im latenten Raum der Netzwerkmodelle, wofür das bereits erwähnte CLIP-Modell der Firma OpenAI[26] zur Verfügung steht. Somit

25 Die Zitate werden hier anonymisiert wiedergegeben, sie sind Teil der durch Dominik Bönisch durchgeführten Expert*inneninterviews. Dominik Bönisch, Francis Hunger: »From Keras Import Curating«: Eine empirische Erhebung zur Übertragung von kuratorischer Praxis auf maschinelle Lernmodelle. Training the Archive – Working Paper 7, Aachen/Dortmund 2023. https://doi.org/10.5281/zenodo.10012483.

26 Alec Radford u.a.: Learning Transferable Visual Models From Natural Language Supervision, in: arXiv, 26.2.2021, http://arxiv.org/abs/2103.00020. Zu Datenbias in CLIP, siehe Agarwal u.a.: Evaluating CLIP: Towards Characterization of Broader

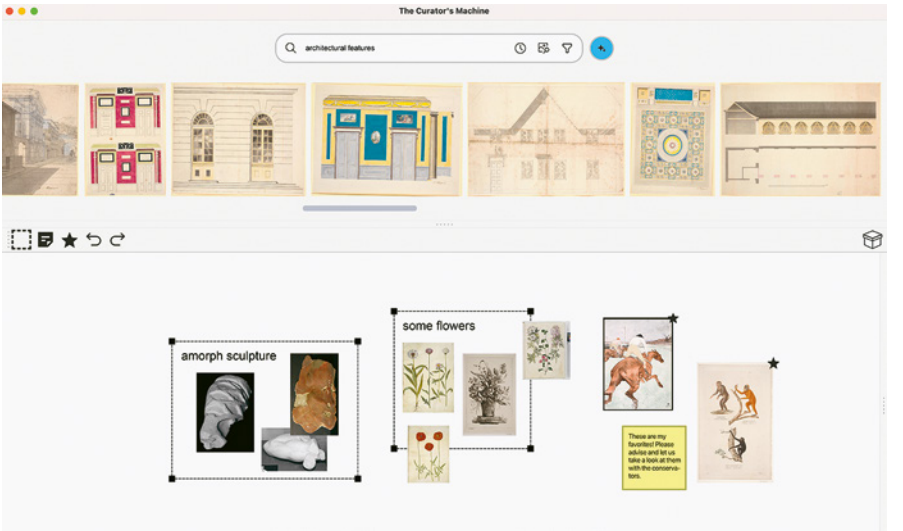

Abb. 4: Mockup des Interfaces der »Curator's Machine«, Stand März 2023

lassen sich Bildmengen nach Phrasen und Stichworten durchsuchen, ohne bereits vorab eine Struktur der Daten kennen zu müssen (zum Beispiel die möglichen Eingabefelder und Eingabewerte, wie es in Datenbanken der Fall wäre). Vielmehr können die Suchenden assoziative Begriffe verwenden, da die Bild-Text-Annäherung des CLIP-Modells korrelativ erfolgt, wie weiter unten ausgeführt wird.

Als dritter, finaler Prototyp wurde eine interpretative Suchmaschine entwickelt, die auf den vorherigen Entwicklungen aufbaute (Abb. 4). Zum einen gibt es nun eine generelle Suche, die mittels der CLIP-Architektur Text- und Bildembeddings ermöglicht und dadurch textbasierte Suchen unterstützt. Es können Phrasen gesucht werden wie ›a portrait with a smiling face‹, ›abstract art with the color beige‹ oder auch ein Medium wie ›painting‹, ohne dass diese Bezeichnungen in den Metadaten hinterlegt sein müssen. Die Beispiele zeigen auch, worin sich der Prototyp III von den Vorgängern unterscheidet: Durch das korrelative CLIP-Embedding können im Unterschied zu ImageNet-trainierten Netzwerken auch generellere beziehungsweise unscharfe Konzepte, wie ›abstract art‹ oder ›love‹ gefunden werden.[27]

Capabilities and Downstream Implications, in: arXiv, 5 Aug 2021, https://arxiv.org/pdf/2108.02818.pdf.

27 Diese Konzepte unterliegen bedauerlichen, aber auch erwartbaren, Einschränkun-

Zum anderen können die Kurator*innen Suchergebnisse verwerfen oder bestätigen, indem sie beginnen, Bilder zu gruppieren. Inspiriert wurde dies von der Vorgehensweise einiger interviewter Kurator*innen, Kunstwerke in kleinem Format auszudrucken und auf einem skizzierten Raumplan zu verteilen, bis sich eine Auswahl und Anordnung ergibt. Die Gruppierungsfunktion von Objekten innerhalb des Prototyp III entspricht in etwa dem Modus des Kuratierens im Sinne eines Zusammenstellens in Haupt- und Unterthemen. Dabei rekonfiguriert sich das aktive Embedding anhand der Interaktionen der Kurator*innen und passt die Suchergebnisse in Echtzeit an die manuellen Sortierungen an.

Die Herausforderungen für die abschließende Entwicklung waren neben a) dem angepassten Machine-Learning-Verfahren, b) ein funktionales Interface, c) die Verkürzung des Embedding-Updates in Echtzeit-Loops und d) die Vernetzung multimodaler Zusammenhänge mittels sogenannter ›Localized Latent Updates‹. Letzteres ist ein neues Verfahren, welches durch den Kooperationspartner, die RWTH Aachen University, entwickelt wurde. Es setzt auf das vortrainierte CLIP-Modell spezifische Adapter auf,[28] da CLIP im Unterschied zu anderen gewichteten Netzen aufgrund seiner Größe (35 Millionen Parameter) nicht ohne weiteres nachtrainiert werden kann. Die Entscheidungen der Kurator*innen über den Ein- oder Ausschluss und das Gruppieren von Werken wurden in diesem neuen Adapter-Verfahren mit dem CLIP-Embedding lokal für die jeweilige Sitzung verknüpft.

Ein Zwischenfazit: Der Anspruch an die »Curator's Machine« bleibt, auch wenn es der nunmehr eingeführte Name suggerieren mag, keiner der Vollautomatisierung. Vielmehr soll die Anwendung dabei helfen, große Datenmengen in Kunstmuseen besser zugänglich zu machen. Durch eine systematische Aufbereitung und Visualisierung von Zusammenhängen und Korrelationen zwischen Kunstwerken können Kurator*innen ein neues nützliches Instrument einsetzen. Der große Vorteil der »Curator's Machine« ist, dass sich die Vorschläge der Nutzer*in anpassen, da die Interaktionen mit der Maschine durch den Feedback-Loop der Localized Latent Updates rückgekoppelt ist.

gen: So findet der CLIP-Embedding-Space zur Suchphrase ›two people in love‹ hauptsächlich heteronormative Bilder und interessanterweise auch viele mit Text, die Sinnsprüche zum Thema ›Liebe‹ enthalten. Dies lässt sich auch anhand der Trainingsdaten nachvollziehen, die unter dem folgenden Link durchsucht werden können: https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn5.laion.ai&index=laion5B&useMclip=false&query=two+people+in+love (zugegriffen am 4.12.2022).

28 Moritz Ibing, Isaak Lim und Leif Kobbelt: Localized Latent Updates for Fine-Tuning Vision-Language Models, in: arXiv, 13.12.2022, http://arxiv.org/abs/2212.06556.

Die Kurator*innen erhalten somit ein Expert*innenwerkzeug. Verbunden damit ist die Hoffnung auf eine erweiterte Auseinandersetzung mit der eigenen musealen Sammlung im Sinne eines Wieder- und Neuentdeckens von Kunst jenseits starrer Datenbankabfragen und abseits der Suche nach bereits bekannten Objekten.

Korrelative, gewichtete Netze und strukturierte Daten in Sammlungen

Im Anschluss sind einige grundsätzliche Überlegungen angebracht, die über die Anwendungspraktiken der »Curator's Machine« hinausgehen. Zunächst fällt auf, dass wir es mit zwei verschiedenen Arten der Datenverarbeitung zu tun haben: Erstens handelt es sich bei den Sammlungsdaten um strukturierte Daten, die in Form von Datenbanken prozessiert und gespeichert werden. Diese Daten zeichnen sich dadurch aus, dass sie in feste Einteilungen eingetragen werden, die durch Formularansichten und Tabellen in Erscheinung treten. Die Notwendigkeit der Struktur lässt sich sogar so weit zuspitzen, dass man sagen kann, dass Daten nur dann zu solchen werden, wenn sie sich in die vorgegebene Ordnung eintragen lassen. Der Bedeutungsgehalt von Daten muss sich mit der Beschreibung der entsprechenden Eintragsfelder decken. Die Information befindet sich hier in Formation.[29]

Dies wird von den Nutzer*innen häufig als einschränkend empfunden. Als problematisch wird beispielsweise wahrgenommen, wenn für bestimmte Informationen kein passendes Feld vorhanden ist. Denn bei relationalen Datenbanken ist das Informationsmodell ausschlaggebend, also die Entscheidung darüber, was in sie aufgenommen werden kann und was nicht – d.h., welche Felder oder Spalten mit Daten gefüllt werden können und welche möglicherweise überhaupt nicht als Kategorien vorhanden sind. Oft wird das Informationsmodell zu Beginn der Nutzung einer Datenbank erstellt, und im Gebrauch zeigen sich andere Bedürfnisse, oder es entstehen neue. Denn im Unterschied zur Verwendung individualisierbarer Tabellenkalkulationen, in denen die User*innen problemlos weitere Spalten hinzufügen können, sind Datenbanken durch andere infrastrukturelle Gebrauchsweisen gekennzeichnet. Sie sind weniger stark personalisiert, sondern dienen oft der Arbeit in übergreifenden infrastrukturellen Situationen, wo sie lokale und globale

29 Vgl. Markus Krajewski: In Formation – Aufstieg und Fall der Tabelle als Paradigma der Datenverarbeitung, in: Nach Feierabend: Zürcher Jahrbuch für Wissensgeschichte – Datenbanken, Zürich 2007, S. 37–55.

Bedürfnisse miteinander vermitteln müssen. Daher können die individuellen Wünsche einzelner Nutzer*innen erst Eingang in eine globale Datenbankanwendung und somit in das Informationsmodell finden, nachdem ein Aushandlungsprozess mit allen betroffenen Akteur*innen stattgefunden hat.

Es ist zu vermerken, wie das Informationsmodell eine kategoriale und strukturierte Repräsentation der Wirklichkeit konfiguriert. So argumentiert der Informatiker Geoffrey Bowker: »[C]ategories are central to being in the world.«[30] In *Sorting Things Out* definieren die Soziologin Susan Star und Geoffrey Bowker Klassifikationen als »spatial, temporal, or spatio-temporal segmentation of the world«[31] mit dem Ziel, (bürokratische) Akte zu vollziehen oder Wissen zu produzieren. Die Medienphilosophin Sybille Krämer bezeichnet Tabellen und Diagramme als zweidimensionale »Werkzeuge des Denkens«, welche In-Formationen erzeugen, indem sie auf Basis von Realität Daten produzieren und in der Fläche verteilen. Diese Daten können wiederum diagrammatisch erschlossen werden.[32] Der Medienwissenschaftler Markus Krajewski verweist auf die Formatierung, welche gleichzeitig eine In-Formatisierung darstelle: »Hier [im Formular – F.H./D.B.] gibt das Format der Daten, ihre Standardisierung und jeweilige tabellarische Anordnung zueinander, den Ausschlag für eine korrekte kategoriale Erfassung und anschließende Verarbeitung.«[33] Das Informationsmodell in (relationalen) Datenbanken basiert auf dieser In-Formatisierung von Wirklichkeit und es erzeugt Ein- und Ausschlüsse.

Strukturen oder Grammatiken bestimmen unser Denken und unsere Denkwerkzeuge wesentlich mit. Es kann folglich als Einschränkung empfunden werden, wenn für eine Suche bestimmte Abfragen in starre Felder eingetragen werden müssen, zum Beispiel der Titel in das Feld ›Titel‹ und Autor*innen in das Feld ›Autor‹. In derartigen Eingabemasken ist (dia-)grammatisch klar und explizit, nach welchen Kategorien die Datenbank und die Inhalte der

30 Geoffrey C. Bowker: The Theory/Data Thing Commentary, in: International Journal of Communication, 8, 2014, S. 1795–1799, hier S. 1797.

31 Geoffrey C. Bowker und Susan Leigh Star: Sorting Things Out – Classification and its Consequences, Cambridge, MA 1999, S. 10.

32 Sybille Krämer: Zwischen Anschauung und Denken – Zur epistemologischen Bedeutung des Graphismus, in: Was sich nicht sagen lässt. Das Nicht-Begriffliche in Wissenschaft, Kunst und Religion, hg. von Joachim Bromand, Berlin 2010, S. 173–192, hier S. 182; Id.: Notationen Schemata und Diagramme – ›Räumlichkeit‹ als Darstellungsprinzip, in: Notationen und choreographisches Denken, hg. von Gabriele Brandstetter, Franck Hofmann und Kirsten Maar, Freiburg i.Br./ Berlin/ Wien 2010, S. 29–45, hier S. 40, 42.

33 M. Krajewski (wie Anm. 29), hier S. 37.

Datenbankstruktur geordnet sind. Durch Einsichtnahme der Felder und über deren Bezeichnungen kann zudem geschlussfolgert werden, welche Ausschlüsse stattgefunden haben.

Wie gestaltete sich aber die Maßgabe struktureller Daten in den Prototypen der »Curator's Machine«, welche auf gewichteten Netzwerken beruhen? Laut dem Medienphilosophen und Literaturwissenschaftler Hannes Bajohr folgen diese einer völlig anderen Logik als relationale Datenbanken, wie er unter Verweis auf Lev Manovichs *The Language of New Media*[34] schreibt: »When there are no explicitly encoded items anymore that can be accessed individually, but only statistical dependencies that are distributed throughout the system, we are confronted not with a database logic but with something else entirely.«[35] Was ist dieses durch Bajohr aufgerufene »something else entirely« im Kontext der »Curator's Machine«? Zunächst ist aufgrund der eingesetzten Algorithmen zwischen den ersten beiden und dem dritten Prototypen zu unterscheiden:

1) Die ersten zwei Prototypen basierten auf ›supervised‹ vortrainierten Netzwerken, deren Gewichte mithilfe der Bild-Datensammlung ImageNet[36] per Backpropagation[37] eingeprägt wurden. Die in ImageNet klassifizierten Trainingsbilder sind manuell annotiert und zwar entlang der Ontologie von »WordNet«. Diese Struktur ist bereits deutlich flexibler und informationstechnisch unschärfer, als dies in relationalen Datenbanken der Fall ist. In vortrainierten Netzen wie Inception, ResNet oder VGG19 lassen sich zwar durchaus strukturgebende Relationen identifizieren, insbesondere weil sie in der in den 1980er-Jahren entwickelten WordNet-Ontologie verwurzelt sind. Jedoch gehen aufgrund vielfacher Reduktionen und Extrapolationen,

34 Lev Manovich: The Language of New Media, Cambridge, MA 2001.

35 Hannes Bajohr: Algorithmic Empathy: Toward a Critique of Aesthetic AI, in: Configurations 30/2 (3/2022), S. 203–231, hier S. 225.

36 Fei-Fei Li u.a. (wie Anm. 14); G.E. Hinton, A. Krizhevsky und I. Sutskever (wie Anm. 14); Fei-Fei Li u.a. (wie Anm. 17).

37 Backpropagation ist ein algorithmischer Prozess, der die Gewichtungen innerhalb gewichteter Netzwerke neu einstellt. Der Ausgabewert eines gewichteten Netzwerkes ermittelt anhand von Testbildern, deren Bedeutung bekannt ist, einen Fehlerquotienten. Dieser wird rückwärts durch die Schichten gewichteter Netzwerke propagiert und führt zu einer algorithmischen Überarbeitung des Netzwerkes. Das Prinzip wurde für heutige gewichtete Netze grundsätzlich beschrieben in David E. Rumelhart, Geoffrey E. Hinton und Ronald J. Williams: Learning representations by back-propagating errors, in: Nature 323/6088 (10.1986), S. 533–536, und Y. LeCun u.a.: Backpropagation Applied to Handwritten Zip Code Recognition, in: Neural Computation 1/4 (12.1989), S. 541–551.

die innerhalb der vielgeschichteten gewichteten Netzwerke vorgenommen werden, die unmittelbaren Relationen zwischen Bild- und Metadaten verloren. Grundsätzlich gilt: Die Bilder werden nicht anhand ihres Bildinhaltes als Bild-Konstruktionen interpretiert, wie Menschen dies tun würden, sondern als statistische Korrelationen, bei denen ein bestimmter Begriff, die Klassifikation, mit dem gehäuften Auftreten bestimmter Pixelformationen assoziiert und im Rahmen des Trainings in das Netzwerk eingespeichert wurde. Konzepte werden von diesen korrelativen Netzwerken nicht ›erkannt‹ und es wird keine ikonographische Deutung betrieben. Es wird also nichts ›verstanden‹ oder ›gelesen‹. Allenfalls speichern gewichtete Netzwerke Kombinationen visueller Repräsentationen, die Features, ab. Folglich sind es auch nicht optische Merkmale, sondern visuelle, d.h. am Stil und an Texturen orientierte computerisierte Korrelationen, mit denen die Nutzer*innen umgehen.[38] Wichtig ist an ImageNet-trainierten Netzwerken, dass die Kategorisierung der Trainingsbilder ursprünglich durch Menschen vorgenommen wurde, die als ›Clickworker*innen‹ intentional tausende Trainingsbilder gesichtet und mit einem Label versehen haben.

2) CLIP, welches »Training the Archive« im dritten Prototyp einsetzte, besteht aus noch größeren Datenmengen als ImageNet. Warum dies für die »Curator's Machine« von besonderem Interesse war, soll im Folgenden betrachtet werden.

Die Annotationen von Trainingsbildern in ImageNet erfolgten durch Clickworker*innen, die für diese Aufgabe eigens angeheuert wurden. Die Annotationen in CLIP hingegen entstanden zwar auch durch Menschenhand, allerdings völlig unabhängig – durch die Nutzer*innen von Internetangeboten, die ihre Fotos online gestellt und mit einer Bildbeschreibung versehen haben. Die Bild-URLs und die Bildbeschreibungen wurden durch OpenAI ungefragt und unbezahlt automatisiert extrahiert. Die genaue Datengrundlage ist bisher unveröffentlicht und daher intransparent.[39] OpenAI

38 Roland Meyer: Im Bildraum von Big Data. Unwahrscheinliche und unvorhergesehene Suchkommandos: Über Dall-E 2, in: cargo 55, 2022, S. 50–53, hier S. 53.

39 In der Model-Card zu CLIP sind die Aussagen zum Datensatz recht allgemein gehalten: »a combination of crawling a handful of websites and using commonly-used pre-existing image datasets such as YFCC100M«, https://github.com/openai/CLIP/blob/main/model-card.md (zugegriffen am 4.12.2022). Zur Problematik des Scrapings für multimodale vortrainierte Netzwerke siehe auch Andy Baio: AI Data Laundering – How Academic and Nonprofit Researchers Shield Tech Companies from Accountability, in: Waxy.org, Blog, 30.9.2022, https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/ (zugegriffen am 4.12.2022). Eine kritische

nutzt diesen gescrapten Datensatz mit dem Ziel, ein größtmögliches Modell vorzutrainieren, welches breite Wissensdomänen abzudecken vermag. Eine Vorstrukturierung, wie beispielsweise durch WordNet, ist hier nicht beabsichtigt. Der Vorteil von CLIP im Vergleich zur Klassifizierung mit ImageNet liegt darin, dass nicht einzelne Worte einem Bild zugeschrieben werden, sondern eine ganze Textphrase, beispielsweise die Wortfolge ›Foto einer niedlichen Katze‹. Dies geschieht anhand riesiger Trainingsdatensätze, so dass die statistische Wahrscheinlichkeit einer richtigen Text-Bild-Korrelation möglichst hoch ist. Doch geht CLIP mittels des Contrastive Language – Image Pre-training noch darüber hinaus. Was bedeutet ›Contrastive‹ in diesem Zusammenhang? Die Wortfolgen werden im Zuge des Trainings untereinander kontrastiert, d.h., es wird errechnet, wie unterschiedlich sie zu anderen Wortfolgen sind.[40] Dadurch entstehen für jedes Text-Bild-Paar Abstände zu anderen Paaren. Einander ähnliche Paare sammeln sich im mehrdimensionalen, latenten Raum zu Clustern. In der Folge können diese Paare sowohl über eine Bildsuche als auch über eine Textphrasensuche abgerufen werden. Der Bildwissenschaftler Roland Meyer beschreibt diese Suchanfragen als »Fahndungsaufrufe, zu denen die Software aus dem latenten Raum möglicher Bilder die aus ihrer Sicht passendsten Entsprechungen hervorbringt.«[41] Tests mit CLIP zeigten, dass es viele unterschiedliche Wissensdomänen enthält, sofern sie im Internet gut vertreten sind. Es performt ähnlich gut oder besser als zum Beispiel die Netzwerke ResNet auf Test-Datensätzen wie ›Food 101‹ oder ›Stanford Cars‹. Umgekehrt bedeutet dies, dass beispielsweise die Detektion von Tumoren oder von Satellitenbildern – also Bildern, die vergleichsweise seltener im Internet mit Bildunterschriften vertreten sind – in CLIP schlechter ausfällt als in spezifisch darauf vortrainierten ResNet-50-Netzen mit dem ›EuroSAT‹- oder dem ›PatchCamelyon‹-Datensatz.[42]

Tests, wie gut CLIP in der Lage ist, Kunst zu klassifizieren, liegen bisher begrenzt vor, doch scheinen CLIP-Ansätze ähnlich gut zu performen wie spezifisch trainierte gewichtete Netze. Die Datenwissenschaftler Marcos Conde und Kerem Turgutlu schlagen vor, dafür die Metadaten von Bildern, welche aus einzelnen Stichworten bestehen, für die CLIP-Verarbeitung in Phrasen zu verwandeln, um bessere Suchergebnisse zu erzielen, zum Beispiel

Diskussion der politischen Ökonomie von Transformer-Netzwerken unternehmen Dieuwertje Luitse und Wiebke Denkena: The great Transformer: Examining the role of large language models in the political economy of AI, in: Big Data & Society, 8/2, 07.2021.

40 A. Radford u.a. (wie Anm. 26), hier S. 2.

41 R. Meyer (wie Anm. 38), hier S. 52.

42 A. Radford u.a. (wie Anm. 26), S. 8.

die Eigenschaften ›Land: Japan; Medium: Paper‹ in die Phrase ›artwork from Japan, made of paper‹.[43]

Wie spiegelt sich das Konzept ›Kunst‹ in CLIP wider? Unabhängig von den in Conde und Tugutlu genannten Metriken ist davon auszugehen, dass zahlreiche Kunstabbildungen durch Lai*innen im Internet zur Verfügung gestellt wurden. Es ist zu vermuten, dass ›berühmte‹ Bildbestände stärker vertreten sind, da diese nicht nur durch das Marketing der Museen, sondern auch durch die Fotografien der Nutzer*innen in Ausstellungen massiv zirkulieren. Des Weiteren haben renommierte Kunstmuseen ihre Datenbestände im Zuge von Google Arts & Culture, dem Europeana-Projekt, der Wikipedia GLAM (Galleries, Libraries, Archives, and Museums) und weiteren Initiativen digitalisiert, mit Expert*innenbeschreibungen versehen und online gestellt.[44] Der urheberrechtsfreie Teil dieser Sammlungen gehört zur Wikimedia Commons und wird darüber distribuiert.

Unter den Begriff ›art‹ fallen im englischen Sprachgebrauch allerdings nicht allein die akademischen Künste. ›Art‹ ist viel weiter gefasst und schließt ebenfalls nicht-akademische, amateurhafte oder volkstümliche Kunstproduktion ein, so wie zum Beispiel die Veröffentlichungen auf Deviant-Art oder die Einordnungen von Lai*innen auf der in CLIP stark vertretenen Bildplattform Pinterest. Genauso, darauf wies Roland Meyer hin, sind viele Reproduktionen Bildender Kunst auf nicht-institutionellen, kommerziellen Seiten (Poster-Shops u.ä.) verfügbar. Dies betrifft vor allem Kunstwerke, deren Urheberrechte erloschen sind und die gleichzeitig als populär genug eingeschätzt werden, um kommerziell verwertbar zu sein. Auch deren Bildbeschreibungen, die vor allem auf Search Engine Optimization und Verkaufserlöse ausgerichtet sind, fließen in Transformermodelle wie CLIP ein. Damit verschiebt sich die Bewertungsgrundlage dafür, was Kunst umfasst und wie diese in den Bildbeschreibungen von CLIP erscheint.

Schließlich basiert die Suchleistung von CLIP auf einem Phänomen, welches Hannes Bajohr als die Produktion ›dummer Bedeutung‹ beschreibt: Multimodale Modelle sind demnach »Produzenten eines ersten Grades dummer Bedeutung. Dumm ist sie, weil das Modell zwar latente Korrelationen zwischen Zeichen erfasst, aber immer noch nicht ›weiß‹, welche Sachen diese

43 Marcos V. Conde und Kerem Turgutlu: CLIP-Art: Contrastive Pre-training for Fine-Grained Art Classification, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Juni 2021.

44 Siehe Google Arts & Culture: https://artsandculture.google.com/, Europeana: https://www.europeana.eu/de, Wikipedia Galleries, Libraries, Archives, and Museums: https://outreach.wikimedia.org/wiki/GLAM (zugegriffen am 20.4.2023).

Zeichen eigentlich benennen.«[45] Kurz gesagt: Die Suchergebnisse basieren auf statistischen Korrelationen, ohne konzeptuelles Verständnis von Kunst.

Dummes ›Korrelationswissen‹ unterscheidet sich von jenem Wissen, das durch Expert*innen in Datenbanken abgelegt ist, da es nicht den gleichen Zugangskontrollen und strukturellen Erwägungen unterliegt und da kein explizites Informationsmodell wie in relationalen Datenbanken formuliert werden muss. Fragen der Repräsentation und Nichtrepräsentation von Personen und Personengruppen in den Trainingsdaten sind zudem bisher ungelöst und schreiben rassistische, sexistische und weitere Diskriminierungen in CLIP fort.[46] Aus dieser Sachlage heraus setzte sich »Training the Archive« das Ziel, den finalen Prototypen mit der Open-Source-Implementierung »OpenCLIP« zu betreiben, welche zumindest eine transparente Darstellung der Modellarchitektur aufweist und dadurch die Monopolstellung von OpenAI unterwandert.[47] Der dritte Prototyp der »Curator's Machine« verknüpfte und verschränkte des Weiteren nun beide Ebenen: Das korrelationale Datenkorpus von CLIP beziehungsweise OpenCLIP diente als Suchfilter für die anhand eines Informationsmodells strukturierte Museumsdatensammlung.

Diese Verknüpfung geschah mehrstufig:

1. Das im CLIP-Verfahren vortrainierte Netzwerk verfügt über unkontrollierte Ordnungen, die nicht intentional sind, sondern sich aus im Internet gescrapten Bild-Text-Paaren ergeben. Dadurch entsteht ein Bias, der in der ungleichen Repräsentation von Lebewesen, Objekten und visuell Nicht-Repräsentierbarem begründet liegt. Anders formuliert: Die Auswahl bestimmter algorithmischer Verfahren und der Datengrundlage formt die Wissenskontexte. Daher kommt dem Auswahlprozess von Algorithmen und Datenmengen, den größtenteils Mathematiker*innen und Ingenieur*innen vornehmen, besondere Bedeutung zu. Aufgrund der Rechenintensivität für die Erstellung von Modellen mit Billionen von Parametern erfolgt die momentan als Durchbruch empfundene Auswahl passender Algorithmen, wie beispielsweise Bidirectional Encoder Representations from Transformers (BERT) und Generative Pretrained Transformer (GPT) anhand (rechen-)ökonomischer Erwägungen. In der schwachen inneren Struktur der Trainingsdatensätze liegt gleichzeitig deren Stärke, weil das vortrainierte Netzwerk CLIP es erlaubt, andere Bestände – im Falle von »Training the

45 H. Bajohr (wie Anm. 6), hier S. 74.

46 Fabian Offert und Thao Phan: A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space, in: ArXiv:2211.06323 [Cs], 26.10.2022, http://arxiv.org/abs/2211.06323.

47 Siehe dazu: https://github.com/mlfoundations/open_clip.

Archive« die digitalisierten Museumssammlungen – mit wesentlich offeneren und unschärferen Fragestellungen zu durchsuchen, als dies die Metadaten in relationalen Datenbanken ermöglichen würden, zum Beispiel: ›Malerei mit Pfirsichen, Äpfeln und einem roten Hintergrund‹ oder ›Grafik mit verliebtem Paar‹.

2. Nachdem mit Hilfe von CLIP eine Vorauswahl getroffen wurde, erhalten die Nutzer*innen der »Curator's Machine« die Möglichkeit, einerseits per User-Interface die Ergebnisse zu filtern und andererseits Gruppierungen zusammenzustellen. Die Filterfunktion greift dabei auf Metadaten zurück, also auf jene strukturierten Daten, die Sammlungskurator*innen für jedes Bild entsprechend dem Informationsmodell eingeschrieben haben. So ist es beispielsweise möglich, Suchen nach einer ›Malerei mit Pfirsichen, Äpfeln und einem roten Hintergrund‹ beziehungsweise die Annäherung daran mittels strukturierter Daten wie ›Jahr‹, ›Größe‹ oder ›Format‹ einzugrenzen. Diese Einschränkungen wirken nicht (!) in die Tiefen des Embeddings zurück, sondern finden allein an der Oberfläche des User-Interfaces ihre Anwendung. Sie verändern daher die Zusammenstellung der Suchmenge nicht dauerhaft, sondern nur für den Moment der Abfrage. Sie zeigen, dass strukturierte Daten und die Erschließung unstrukturierter Datenmengen sich überlappen können.
3. Einer der neuen Ansätze der »Curator's Machine« ist die Möglichkeit, zusätzlich zu den in 1. und 2. genannten Punkten die Ergebnismenge per Drag-and-Drop zu Gruppen zusammenzufassen. Das entspricht der von Kurator*innen nachgefragten Auswahlfunktion. Dieses Verfahren kommt vor allem bei großen Datenmengen zum Zuge, denn die manuelle Gruppierung wirkt lokal zurück in den latenten Raum und holt anhand ähnlicher Vektoren weitere, ähnliche Bilder als Vorschläge aus den Tiefen der Sammlung hervor.

Fazit

Die eindeutige Zuordnung von Elementen und Kategorien mittels strukturierter Daten in Datenbanklogiken unterscheidet sich deutlich vom korrelativen Clustering im latenten Raum eines gewichteten Netzwerkes. In letzterem sind die Beziehungen zu Datenstrukturen wesentlich gelockert, beziehungsweise opak. Am Beispiel der »Curator's Machine« wurde gezeigt, wie die strukturiert vorliegenden Daten der Sammlungen in der Software verworfen und dann per Korrelation neu zusammengesetzt werden, um wieder Struktur zu erhalten. Die durch das Informationsmodell von Datenbanken vorgegebene Ordnung wird temporär aufgelöst. Dabei geraten die in Sammlungsdatenbanken encodierten In-Formationen aus der Formation. Multimodale gewichtete Netzwerke erlauben mittels statistischer Verfahren,

wie der K-Nearest-Neighbors oder dem Contrastive Language – Image Pretraining, das computerisierte Prozessieren von Unschärfe.

Diese jeweils verschiedenen Zugriffsweisen auf Daten verkomplizieren sich im Zuge der »Curator's Machine« weiter: Auf die gewichteten Netzwerke künstlicher ›Intelligenz‹ setzt ein User-Interface auf, welches a) menschliche Interaktionen mit den Datensammlungen in die Netzwerke feedbackt und b) erneut strukturierte Informationen aus den Metadaten der Sammlungsobjekte heranzieht, um zusätzliche Ordnungskriterien, wie zum Beispiel Sortierungen nach ›Jahr‹ oder ›Titel‹, zur Verfügung zu stellen. Damit wurde aufgezeigt, wie datenbankbasierte Sammlungen und korrelative gewichtete Netzwerke in der »Curator's Machine« anders ineinandergreifen, als dies bei reinen Datenbankanwendungen der Fall ist.[48]

48 Wir danken Hannes Bajohr und Roland Meyer für Anmerkungen und Kommentare sowie Mailin Haberland und Dr. Nora Riediger für Korrekturen.