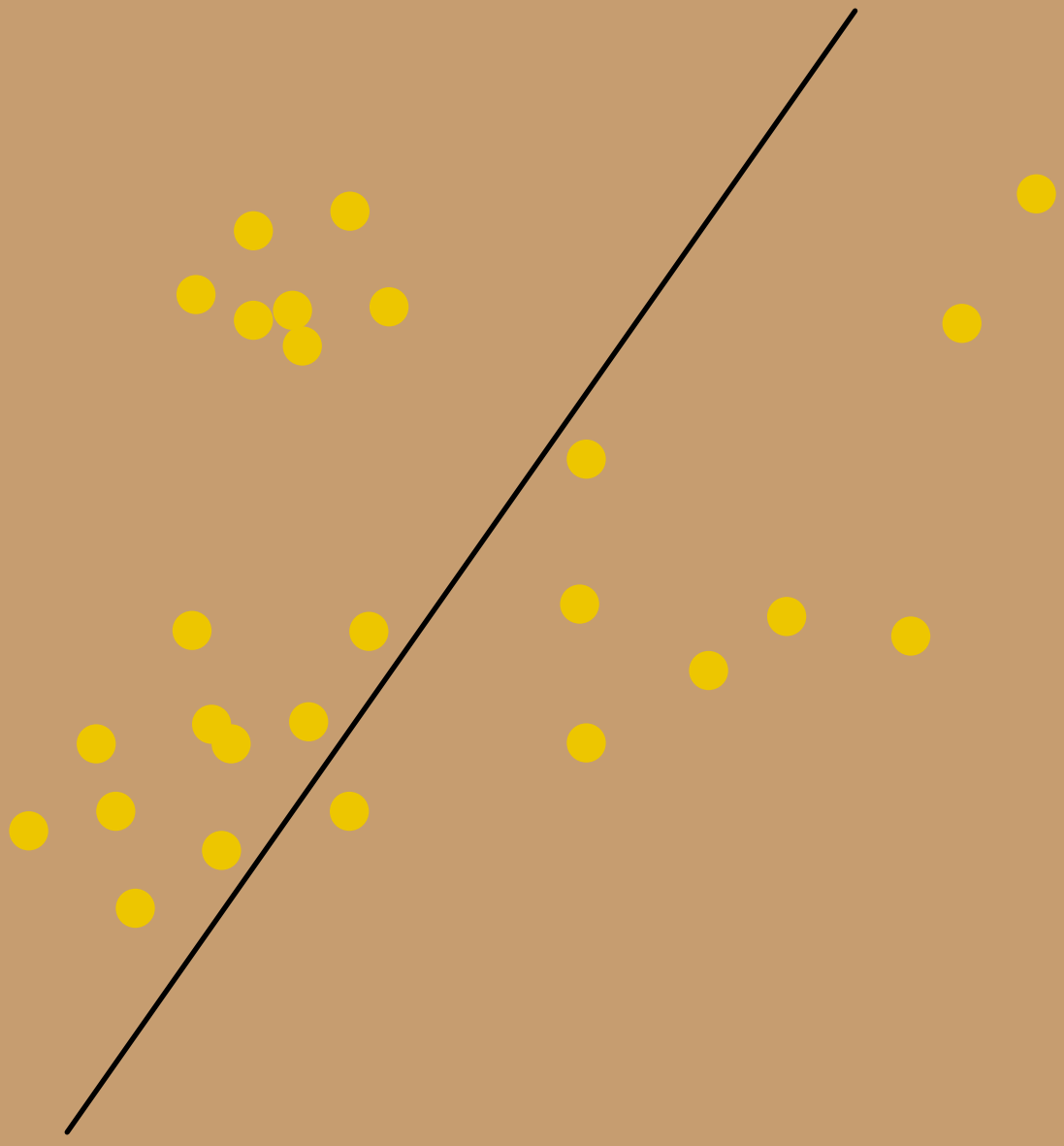
Training the Archive – Working Paper 2

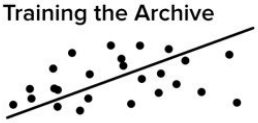


Francis Hunger

„Why so many windows?“ – Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst.

Training the Archive – Working Paper Series
Francis Hunger – „Why so many windows?“ – Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst.
Training the Archive (Hrsg.), Aachen/Dortmund, Mai 2021
DOI: 10.5281/zenodo.4742621

Training the Archive



Ludwig Forum für Internationale Kunst
Jülicher Str. 97-109, 52070 Aachen
http://ludwigforum.de/

Ludwig Forum für Internationale Kunst Aachen

Hartware MedienKunstVerein
Büro: Hoher Wall 15, 44137 Dortmund
www.hmkv.de

HMKV Hartware MedienKunstVerein

Gefördert im Programm Kultur Digital der Kulturstiftung des Bundes

Ku/tur Digita/ KULTURSTIFTUNG DES BUNDES

Gefördert von der Beauftragten der Bundesregierung für Kultur und Medien

Die Beauftragte der Bundesregierung für Kultur und Medien

Working Paper 2: „Why so many windows?“ – Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst.

1. Why so many windows? 2. Why so many tables? 3. Why so many cushions? 4. Why so many close ups? 5. Why so many elephants? 6. Why so many cats? 7. Why so many things related to skate? 8. Why so many computers? 9. Why so many umbrellas? 10. Why so many “Sky plc - company tv cables”? (Pereira und Moreschi 2020, 22)

Abstract

Im Feld der automatisierten Bilderkennung, der sogenannten Computer Vision beziehungsweise Künstlichen ‚Intelligenz‘, hat die Bilddatensammlung ImageNet eine zentrale Rolle als Trainingsdatensatz inne. Für das Forschungsprojekt Training The Archive, welches Methoden der Digital Humanities für das Kuratieren von Kunst verfügbar machen soll, wird erörtert, in welchem Maße ImageNet den Software-Prototypen The Curator’s Machine beeinflusst. The Curator’s Machine soll Zusammenhänge und Verbindungen zwischen Kunstwerken für Kurator*innen erschließen. Es ist bekannt, dass die Trainingsdatensätze ‚neuronaler‘ Netze für Verzerrungen (Bias) in den Ergebnissen sorgen. Wie das in zeitgenössischen Bilderwelten verankterte ImageNet auf zeitgenössische und historischer Kunstwerke einwirkt, erläutert der Text, indem er 1.) die Abwesenheit der Klassifikation ‚Kunst‘ in ImageNet untersucht, 2.) die fehlende Historizität von ImageNet hinterfragt und 3.) das Verhältnis von Textur und Umriss in automatisierter Bilderkennung mit ImageNet diskutiert. Diese Untersuchung ist wichtig für die genealogische, kunsthistorische und programmiertechnische Verwendung von ImageNet in den Feldern des Kuratierens, der Kunstgeschichte, der Kunstwissenschaften und der Digital Humanities.

---

1 Hinführung

Künstliche ‚Intelligenz‘, Machine Learning, Computer Vision sind Automatisierungspraktiken, welche neues Wissen versprechen. Was geschieht, wenn gerahmte Kunst als Fernseher oder Fenster detektiert wird? Was bedeutet es, wenn der Faltenwurf in gotischen Gemälden und Skulpturen anhand des Ikea-Produktkataloges eingeordnet wird? Was passiert, wenn ein

Gemälde Lucas Cranchs des Jüngeren nicht anhand der Umrisse, sondern der Textur in ‚neuronalen' bzw. gewichteten Netzen[1] verarbeitet wird?

Angesiedelt im Feld der Computer Vision[2] untersucht Training the Archive „ein maschinengestütztes, exploratives (Wieder-)Entdecken von Verknüpfungen innerhalb musealer Sammlungen" (Bönisch 2021, 1). Das vorliegende Working Paper 2 bezieht sich grundsätzlich auf das Working Paper 1 von Dominik Bönisch *The Curator's Machine. Clustering von musealen Sammlungsdaten durch Annotieren verdeckter Beziehungsmuster zwischen Kunstwerken* (ebd.). Ausgehend von einem konkreten Software-Prototyp, ‚The Curator's Machine', welcher durch Training the Archive entwickelt wurde und als Open Source dokumentiert ist,[3] werden im Folgenden Fragen aufgeworfen, die für das Zusammenspiel der Felder Kunst, Kuratieren, Kunstgeschichte, Digital Humanities und Computer Vision relevant sind. Grundkenntnisse der Funktionsweisen gewichteter Netze und automatisierter Bilderkennung werden für das Textverständnis vorausgesetzt.[4]

Zur Struktur des Working Papers: Im ersten Abschnitt werden die zu verarbeitenden Bilddaten als ‚operative' Bilder gekennzeichnet. Während repräsentative Bilder auf die Bildinhalte abzielen, die explizit von Menschen für Menschen gemacht sind, existieren operative Bilder, um durch Maschinen verarbeitet werden zu können. Anschließend daran wird die Operationalisierung von Bildern in jenen großen Bildsammlungen untersucht, welche die künstlichen, gewichteten Netze, wie zum Beispiel VGG16 oder InceptionV3, trainieren. Ein nächster Abschnitt widmet sich der Abwesenheit von Kunst in der Bilddatensammlung ImageNet, welche zu den Trainingsstandards heutiger Computer Vision zählt. Der darauf folgende Abschnitt diskutiert die Zeitgenossenschaft operationaler Bilder in ImageNet im Verhältnis zur Historizität von Malerei, Grafik und Skulptur in Museumssammlungen. Darauf folgt eine Diskussion von Textur und Umriss. Wie korrespondieren Textur und Umriss mit den zeitgenössischen Trainingsbildern in ImageNet und dem historischen Bildmaterial? Der Schwerpunkt des vorliegenden Working Papers soll dabei nicht, auf den andernorts bereits erörterten Problemen der Klassifikation, wie zum Beispiel „Bias" liegen (Noble 2018; Crawford und Paglen 2019). Vielmehr sind hier Bild-‚immanente' Problematiken zu verfolgen, zum Beispiel die Frage, wie sich Pre-Processing, Bildformate und historische Formensprache in den späteren Zuordnungsleistungen trainierter gewichteter Netzwerke auswirken.

Warum erfolgt diese Fokussierung auf die Bilddatensammlung ImageNet? ImageNet ist derzeit äußerst weit verbreitet und wird in zahlreichen Forschungsansätzen als Benchmark für die Effektivität gewichteter Netze verwendet. Zwar treten momentan weitere

---

[1] Um den Diskurs zu de-antropomorphisieren, wird hier anstelle von ‚künstlichen neuronalen Netzen' (KNN) der Terminus ‚gewichtete Netze' verwendet. Die Netze, welche in den 1960er Jahren ursprünglich als Neuronen-ähnlich konzipiert wurden (Rosenblatt 1957), sind gekennzeichnet durch gewichtete Knoten (Nodes), welche jedoch nach heutigem wissenschaftlichen Stand mit der Funktion von Neuronen im menschlichen Körper *nicht* übereinstimmen (Cardon, Cointet und Mazieres 2018, 8).

[2] Die Verwendung des Begriffes Computer Vision ist bis in die 1960er Jahre zurückzuführen. Im Unterschied zu Digital Image Processing, welches sich auf das automatisierte Verarbeiten zweidimensionaler Bilder bezieht und zum Beispiel Fragen der Buchstabenerkennung (OCR) adressiert, soll Computer Vision komplexe Bildzusammenhänge automatisieren können, zum Beispiel die Bewegungen und Interaktionen von Objekten in Bildern detektieren. Dieser Ansatz zielt letztlich auf Entscheidungssysteme ab.

[3] https://github.com/DominikBoenisch/Training-the-Archive/.

[4] Empfehlenswerte Einführungen aus geisteswissenschaftlicher Perspektive sind: *How the machine 'thinks' – Understanding Opacity in Machine Learning Algorithms* (Burrell 2016, 5–7) und *How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence* (Pasquinelli 2019, 4–14).

Bildsammlungen von ähnlichem oder größerem Umfang auf, wie zum Beispiel der Open Source Datensatz Open Images von Google, doch wird auch ImageNet weiterentwickelt und zehrt von dem Vorteil, Pionier im Forschungsfeld der Computer Vision gewesen zu sein. ImageNet umfasst als „Canonical Training Set" (Crawford und Paglen 2019) 14 Millionen mit Labels annotierte Bilder, die ihren Inhalt beschreiben sollen (Li u. a. 2009; Krizhevsky, Sutskever und Hinton 2012).[5] Das Bildmaterial reicht von Amateur- und professioneller Fotografie, welche größtenteils von der Fotografieplattform Flickr heruntergeladen wurde, bis zu Produkt- und Stock-Fotografie, die von kommerziellen Websites stammt. Eine kritische Auseinandersetzung mit ImageNet befürworten zum Beispiel Offert und Bell: „Ein breiterer kritischer Ansatz bestünde in der Analyse weit verbreiteter Datensätze wie ImageNet, die nicht nur ‚wie vorgefertigt' in den üblichen Klassifikationsszenarien eingesetzt werden. Öfter noch dienen sie dazu, Klassifikatoren vorzutrainieren, welche dann auf einem separaten Datensatz abgestimmt werden, was möglicherweise ImageNet-Verzerrungen in ein völlig separates Klassifizierungsproblem einbringt" (Offert und Bell 2020, 9).[6]

Das vorliegende Paper konzentriert sich auf ImageNet aus einem weiteren Grund. Die im Feld der Science and Technology Studies verankerte Professorin N. Katherine Hayles verweist darauf, inwiefern *Daten* den Rahmen computergestützter Berechnung bestimmen: „[…] das System vermag die Welt nur anhand der Modalitäten kennen, welche durch seinen Konstrukteur vorgeschrieben wurden. Obwohl es mit diesen Daten neue Ergebnisse erzeugen kann, ist der Rahmen des Neuen durch sein Operationsgebiet – die Daten, die seine Welt erschaffen und beschreiben – begrenzt, im Vorhinein festgelegt ohne die Möglichkeit freier Innovation" (Hayles 2005, 137). Für den Prototypen der Curator's Machine waren die Bilddatensammlung ImageNet ausschlaggebend. Dies sind die Daten, die den von Hayles angesprochenen „Rahmen des Neuen" ausmachen.

Gewichtete Netze zielen üblicherweise auf eine Klassifizierung von Objekten innerhalb eines Bildes ab. Doch soll The Curator's Machine nicht Bildinhalte detektieren, sondern Ähnlichkeiten der Bilder, sogenannte Features. Im Zuge von Training the Archive wird daher die abschließende Klassifikationskomponente des gewichteten Netzes abgeschaltet (siehe Abb. 3). Stattdessen werden die bis dahin berechneten Features für jedes Bild aus den Eingangsdaten anhand der mit ImageNet vortrainierten Gewichte für die weitere Verarbeitung abgespeichert.[7] Es folgt daraus die Frage, welchen Einfluss die Bilddatensammlung ImageNet auf die Feature-Extraktion[8] hat.

Bevor dieser Frage nachgegangen werden kann, ist jedoch erst einmal der Status jener Bilder zu klären, von denen die Rede ist. Sind es überhaupt Bilder? Sind es Daten? Inwiefern die Bilder, die durch Computer-Vision-Systeme geschleust werden, ‚andere' Bilder sind, soll im nächsten Abschnitt erkundet werden.

---

[5] Zur Genealogie von ImageNet siehe auch *Excavating AI* (Crawford und Paglen 2019) und *Lines of Sight* (Hanna u. a. 2020).

[6] Übers. durch den Verfasser. Im Sinne einer besseren Textverständlichkeit wurden fremdsprachige Zitate im gesamten Text durch den Verfasser übersetzt. Die Originalzitate sind in den Endnoten aufgeführt.

[7] Die Vorgehensweise ist hier vereinfacht dargestellt. Tatsächlich wurden die Feature verschiedener vortrainierter Netzwerke (InceptionV3, BiT/m-r152×4 und einzelne Layer aus VGG19) für jedes der 42.000 Eingangsbilder aus der SMK-Sammlung extrahiert und dann weiter verarbeitet. Alle diese Netzwerke sind mit ImageNet vortrainiert.

[8] Siehe auch: https://github.com/DominikBoenisch/Training-the-Archive/blob/master/Prototype/2_Feature_Extractor/Feature_Extractor_Keras_Applications.ipynb.

2 Zwischen repräsentativen und operativen Bildern

Die Bilder aus Bildarchiven, welche im Zuge von Training the Archive durch gewichtete Netzwerke prozessiert werden, ändern ihren Status von ‚repräsentativen' zu ‚operativen' Bildern. Repräsentativ bezeichnet im folgenden nicht-operative Bilder, also Bilder, die auf Interpretierbarkeit angelegt sind, die Ideen vermitteln, ganz gleich ob es sich um gegenständliche oder abstrakte Bilder handelt. Operativ sind Bilder dann, wenn sie eine Reihe automatisierter Operationen ermöglichen, zum Beispiel Identifizierung, Kontrolle, Visualisierung, Erkennung (Broeckmann 2016, 128–134). Operative Bilder sind als encodierte Datenmenge eingebunden in Prozessketten, deren vorrangiges Ziel die Automatisierung von wissensbildenden Verfahren ist. Da ihr Zweck nicht Repräsentation, sondern Operation ist, werden operative Bilder nach der Aufnahme oder dem Scan einem Pre-Processing unterworfen – einer Reihe von Bildmanipulationen, deren Zweck es ist, die Datenmenge für die weitere Verarbeitung vorzubereiten. Über die Homogenisierung operativer Bilder gilt es, in den folgenden Absätzen nachzudenken.

Der Programmierer und Künstler Nicolas Malevé arbeitete heraus, wie die fotografische Aufnahme als homogenisierendes Verfahren zum Einsatz kommt. In der Vorbereitungsphase sollen die eingehenden Bilddaten möglichst formatiert und einander vergleichbar sein: „Eine zentrale Rolle kommt der Fotografie zu, die als Nivellierer konzipiert ist, ein Instrument, das Licht automatisch in Pixel nach festen Regeln umwandelt und das gleichzeitig ein Konzept ab-bildet. Fotografie wird als Instrument zur Homogenisierung der visuellen Welt mobilisiert, um das Visuelle in Daten zu verwandeln, wodurch Daten unterschiedlicher Herkunft verglichen und klassifiziert werden können" (Malevé 2020, 6).

Genereller gesprochen: Um Homogenisierung zu erreichen, werden operationale Bilder einem arbeitsintensiven Pre-Processing zugeführt, in dessen Zuge Belichtungsverhältnisse, Farbraum, Verzerrungen, Größenverhältnisse, horizontale Ausrichtung und ähnliche Parameter vereinheitlicht werden sollen (Brownlee 2019; Chaki und Dey 2020). Zusätzlich zu dem Pre-Processing werden die Bilder mit teils automatisierten und teils manuell erstellten Annotationen versehen, sogenannten Metadaten, zum Beispiel Zeit und Ort der Aufnahme, verwendete Apparate, Namen der Bearbeiter*innen, Bearbeitungsstatus.

Die existierenden Bildarchive sind mit fotooptischen Sensoren von Kameras, Scannern und ähnlichen bildgebenden Geräten (z.B. auch Infrarot, UV, MRI oder Radar) aufgenommen worden (vgl. Amat und Casals 1992; Parikka 2021). Unterschiedliche Archive arbeiten dafür mit verschiedensten Kameratechniken und Objektiven, was dazu führt, dass eine Interoperationalität der Bilder zwischen verschiedenen Archiven eingeschränkt ist. Selbst innerhalb ein und desselben Archives ist die Vergleichbarkeit operativer Bilder prekär. Die Vergleichbarkeit ist abhängig von den handwerklichen Fähigkeiten wechselnder Fotograf*innen, den verschiedenen Geräten, die aufgrund von Verschleiß im Laufe der Zeit durch neue Gerätschaften und dazugehörige Software ersetzt werden, sich wandelnden Informationsbedürfnissen, die institutionellen Schwankungen unterliegen, und sich ändernden Standards des Wissensfeldes (z.B. der Kunstgeschichte oder der Digital Humanities). Da The Curator's Machine derzeit noch nicht mit eigenen Datensätzen arbeitet – diese sind in Vorbereitung–, sondern auf durch Dritte zur Verfügung gestellte operationale Bilder zugreift, übernimmt das Projekt dort eingeschriebenen Homogenisierungen.

Diese Homogenisierung der Datensätze korrespondiert mit einer Beobachtung im Zuge des ersten Prototypen The Curator's Machine. Die den fotografierten Kunstwerken beigefügten Farbkeile, die der fotografischen Angleichung dienen, sind für die anschließende Operationalisierung problematisch. Im Zuge des Pre-Processing müssen sie wieder aus dem Bild entfernt werden.

Zusammenfassend ergibt sich eine Abfolge der Bildmanipulationen, welche die Bilder in zunehmenden Maße operational machen. Dies geschieht in einem eigenständigen Schritt im Zuge der grundlegenden digitalen Erfassung einer Sammlung, die somit zur Bilddatensammlung wird:

1. Einrichten des zu fotografierenden Objektes in neutraler Bildumgebung (Licht, Bildhintergrund)
2. Fotografie oder Scan
3. Annotation und Anreicherung mit Metadaten
4. Allgemeine Bildnachbearbeitung mit dem Ziel einer homogenen Bilddatensammlung. Dies kann zum Teil auch das Bereinigen von „Alterungserscheinungen“ beinhalten
5. Langzeitarchivierung und gegebenfalls Publikation der Sammlung

Für den Fall, dass ein Forschungsprojekt wie Training the Archive auf eine derartige Bilddatensammlung zugreift, um diese im Zuge von Computer Vision weiter zu verarbeiten, treten weitere Arbeitsschritte auf:

6. Bereinigung des Bilddatensatzes (z.B. Doubletten entfernen, Farbkeile entfernen, Ränder beschneiden)
7. Formatanpassung der Bilddaten, Beschneiden auf quadratisches Bildformat für die Prozessierung durch gewichtete Netze, wie zum Beispiel InceptionV3, oder VGG 19.

Es wird deutlich, dass bereits der Status der ‚Ursprungs‘-Bilder, welche einem Pre-Processing unterzogen wurden, von mannigfaltigen Formatierungen gekennzeichnet ist.

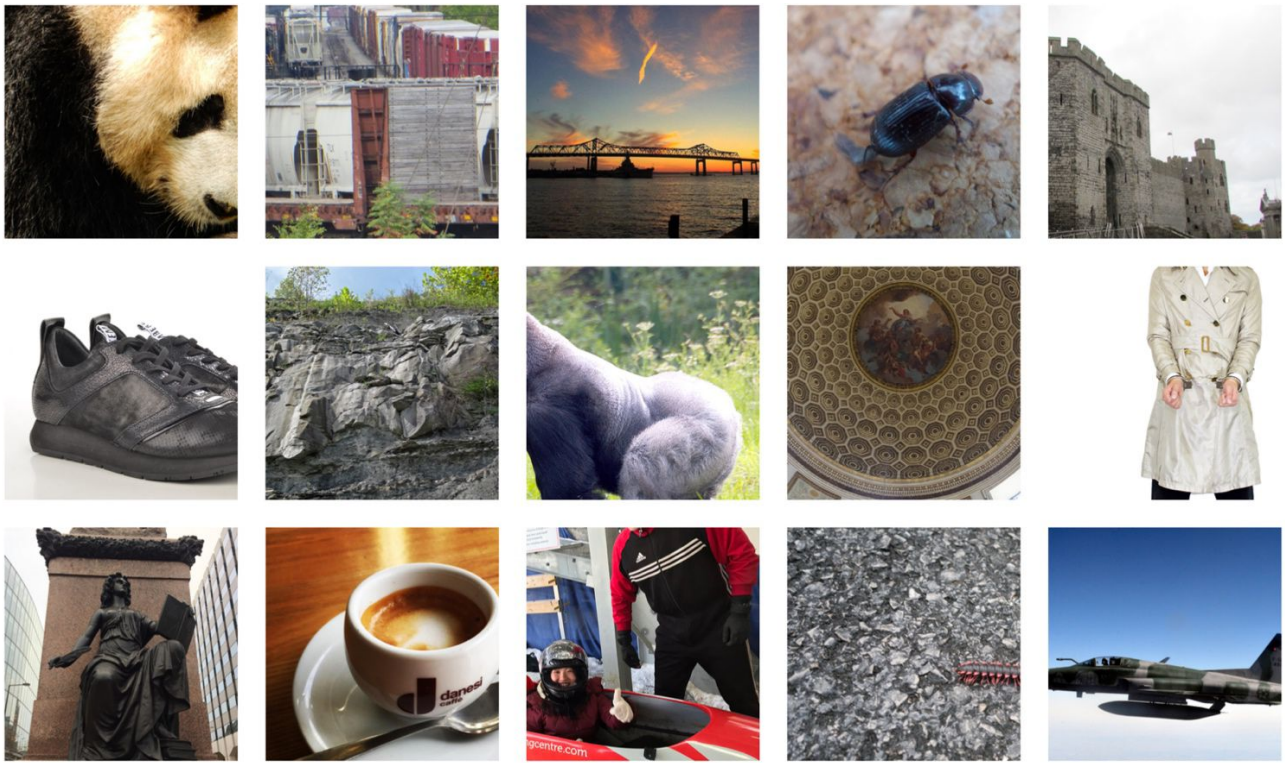


Abb. 1: Operative Bilder 1 – Trainingsdaten aus ImageNet (Autor; das Urheberrecht für die einzelnen Bilder liegt bei den Urheber*innen).

Der Bildbegriff kommt hier an seine Grenzen, handelt es sich doch um Datenmengen, die während des Pre-Processings bereits Berechnungen unterzogen wurden, *bevor* sie überhaupt in die eigentliche Rechenapparatur, das gewichtete Netz, eingeschleust werden: „[…] wir müssen über das Bild – oder sogar den Sensor – als einzelne Einheit hinausblicken und stattdessen verstehen, dass das Bild bestenfalls ein Interface ist (Bratton 2015: 220–6; Andersen und Pold 2018), das eine Art Zugang zu anderen Skalen infrastrukturellen Handelns ermöglicht, welche vielfältige Arten von Wissen über große, dynamische Systeme mobilisieren […]" (Parikka 2021, 203). In Bezug auf gewichtete Netze gilt es, zwischen zwei verschiedenen Formen operativer Bilder zu unterscheiden:

1.) Operative Bilder, welche die gewichteten Netze trainieren (Abb. 1), das heißt den Knoten überhaupt erst Gewichtungen einschreiben. Die gewichteten Netze InceptionV3, BiT/m-r152×4 und VGG19, welche für den ersten Prototypen von The Curator's Machine verwendet wurden, basieren auf dem Training mittels der operativen Bilder von ImageNet.[9] Diese operativen Bilder werden im Folgenden als ‚Trainingsdaten' bezeichnet.

2.) Operative Bilder einer Kunstsammlung, die durch gewichtete Netze klassifiziert werden sollen (Abb. 2). Der erste Prototyp von Training the Archive basierte auf den Eingangsdaten der dänischen Nationalgalerie Statens Museum for Kunst. Sie wurden aufgrund der freien Verfügbarkeit und der „qualitativ hochwertigen Datensätze, zum Beispiel in Bezug auf die vorliegende Bildauflösung, Varianz der Daten und Menge an Metainformationen" (Bönisch 2021, 4) ausgewählt. Im Unterschied zum ersten Punkt sind die Eingangsdaten durch Pre-Processing nachbearbeitete und möglichst vereinheitlichte und idealisierte Abbildungen der Sammlungsgegenstände. Diese operativen Bilder werden im Folgenden als ‚Eingangsdaten' bezeichnet.

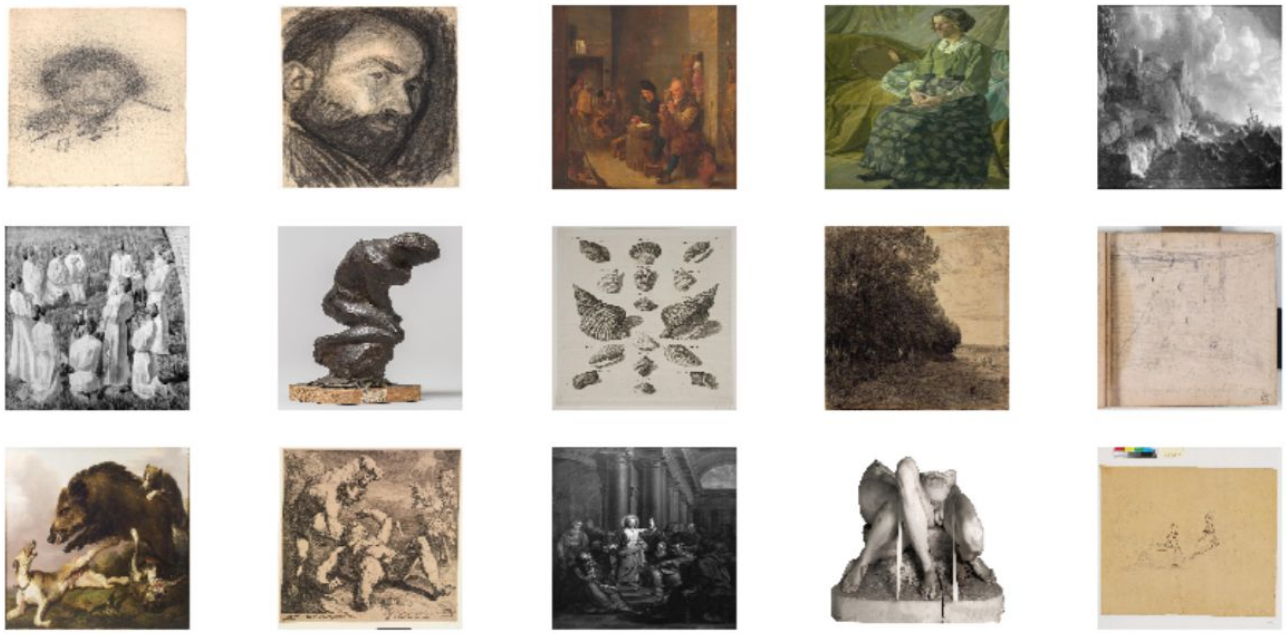


Abb. 2: Operative Bilder 2 – Eingangsdaten der dänischen Nationalgalerie Statens Museum for Kunst, quadratische Bildausschnitte (Bönisch 2021, die Einzelbilder sind in Public Domain).

[9] Da die gewichteten Netze für den Prototypen mittels Keras/Tensorflow initialisiert wurden, ist die dortige Implementierung maßgeblich. Vergleiche: https://keras.io/api/applications/.

Für das Operationalisieren in Computer-Vision-Bibliotheken wie Pytorch oder Tensorflow/Keras haben sich aus rechenökonomischen Gründen Pixelanzahlen von nur 512×512 Pixel (MobileNet V3 Large-M), 299×299 Pixel (InceptionV3), 224×224 Pixel (Resnet50, VGG16, VGG19) und weitere in einem quadratischen Format durchgesetzt.[10]

Als letzte Voraussetzung, bevor die Analyse begonnen werden kann, ist auf einen nicht unwesentlichen Ausschluss hinzuweisen. Dieser Ausschluss bezieht sich darauf, welche künstlerischen Verfahren mit Hilfe von Computer Vision im Zuge von Training the Archive als Eingangsdaten überhaupt erfasst werden. Die beteiligten Institutionen unterscheiden sich hinsichtlich des zur Verfügung stehenden Materials. Während das Ludwig Forum Aachen über eine teildigitalisierte Sammlung aus Malerei, Skulptur und Installationen und Videokunst seit den 1960er Jahren verfügt, definiert sich der Hartware MedienKunstVerein Dortmund, der über keine eigene Sammlung verfügt, allein über seine Ausstellungstätigkeit, welche verschiedenste zeitgenössische Medien sowie konzeptuelle und experimentelle Vorgehensweisen umfasst, jedoch selten Malerei.

Operative Bilder, jenseits der Digitalisierung zweidimensionaler Werke, unterliegen zahlreichen Beschränkungen darin, den künstlerischen Gehalt tatsächlich zu fassen (vgl. Graham und Cook 2010; Grau, Hoth und Wandl-Vogt 2019). Die Eingangsdaten bestehen aus Abbildungen und geben Materialitäten oder konzeptuelle Strategien nicht wieder. Wenn The Curator's Machine mit Mitteln der nicht-semantischen Computer Vision[11] ein Empfehlungs- und Zuordungssystem entwickelt (Bönisch 2021), so läuft das Vorhaben Gefahr, Ausschlüsse der häufig ephemeren, multimedialen oder konzeptuellen künstlerischen Strategien der Gegenwartskunst zu erzeugen.

Die derzeitige Vorgehensweise mit Sammlungsdaten des Statens Museum for Kunst zu arbeiten, ist pragmatisch durch nicht zur Verfügung stehende Digitalisierungsdaten und die große Komplexität des Vorhabens begründet. Doch verstärkt die auf Malerei, Grafik und wenige Skulpturen und Installationen konzentrierte Bilddatensammlung des Statens Museum for Kunst das problematische Normativ von Kunst als zweidimensionales Medium. Es ist daher auch Aufgabe von Training the Archive, Strategien für ephemere, multimediale und konzeptuelle Gegenwartskunst zu finden.

Die hier verwendeten operativen Bilder beziehen sich momentan vorrangig auf zweidimensional abbildbare Werke als Eingangsdaten. Mit dieser Einschränkung gilt es nun zu fragen, was Computer Vision als kuratorisches Werkzeug für einen beschränkten Werkkorpus (an Malerei und Zeichnungen) leisten kann.

---

[10] Die optimalen Pixelwerte werden experimentell ermittelt und haben sich mittlerweile als Standard etabliert. Einer der Gründe scheint darin zu liegen, dass das Bild von der jeweils kleinsten Feature Map eines gewichteten Netzwerks verarbeitbar sein soll und die Input-Daten nicht zu groß sein sollen: „In general, initialized networks with an input size of 224×224px obtained the best results: Xception, InceptionV3, and MobileNet (above 99%); SqueezeNet also obtained competitive results (98.36%) with a smaller input (192×192px)" (Alashhab, Gallego und Lozano 2019, 4).

[11] Die im momentanen Projektzustand von *Training The Archive* verwendeten Verfahren der Computer Vision, vorrangig die gewichteten Netze, operationalisieren die Bilddatenmengen anhand von Mustern und Features, explizit unter Absehen von ihrem semantischen Gehalt.

3 Problemstellungen operativer Bilder und gewichteter Netze

3.1 Die Abwesenheit der Klassifikation „Kunst“ in ImageNet

Aktuelle Untersuchungen haben gezeigt, dass gewichtete Netze teilweise Kunst nicht als solche erkennen. Mit Hilfe der Plattform Wolfram Alpha untersuchte die Künstlerin Rosemary Lee die ersten 100 Treffer für das Stichwort „abstract“ der Bilddatenbank des Metropolitan Museum of Art. Lee stellte fest, dass 98% nicht als Kunst klassifiziert wurden (Lee 2020, 92).[12] Zu ähnlichen Ergebnissen kommen Pereira und Moreschi, und zwar nicht nur in Bezug auf einen Bildkorpus abstrakter Malerei. Computer Vision interpretiere Kunst vor allem als Alltagsgegenstände: „diese Lesarten laden uns dazu ein, Kunstwerke auf eine Art und Weise zu sehen, die von der Idee der Autorenschaft losgelöst ist“ (Pereira und Moreschi 2020, 6).

Von diesen Beobachtungen ausgehend stellt sich die Frage, ob und wie Kunst in den zugrundeliegenden Trainingsdaten der gewichteten Netze vorkommt. Hierzu wurde die heute weit verbreitete Datensammlung operativer Trainingsbilder ImageNet Fall 2011 (Li u. a. 2009; Krizhevsky, Sutskever und Hinton 2012) untersucht. Die Annotationen der in ImageNet gesammelten 15 Millionen operativen Bilder mit 1000 Klassifikationen entstammen einer psycholinguistischen Systematik, welche ab 1985 für die Datenbank Wordnet an der Princeton University erstellt wurde. In Wordnet werden Verben und Substantive (wie z.B. Stuhl, Kind, Kunst) mit IDs versehen und untereinander verlinkt, sodass sich Ketten aus Zugehörigkeiten ergeben, sogenannte Synsets (Synonym Mengen). Die Wordnet-Synsets werden für die Bilddatensammlung ImageNet verwendet, um Bilder zu annotieren und Semantiken zuzuschreiben. Die Konzepte „Kunst“ und „Malerei“ und „Bilderrahmen“ sind in Wordnet vorhanden und müssten prinzipiell auch in ImageNet ansprechbar sein.[13]
Sind in den Trainingsdaten von ImageNet 2011 operative Bilder enthalten, die gesuchten Kategorien betreffen? Meine Untersuchung zeigt: In den ImageNet 2011 Trainingsdaten scheinen unter 15 Millionen Bildern keine vorhanden zu sein, die als „art“ oder „painting“ oder „frame“ gelabelt sind.[14] Das schließt das Vorhandensein von Kunst in ImageNet 2011 nicht aus. Jedoch ist Kunst nicht als solche gelabelt und steht daher Verfahren der Klassifikation in vortrainierten gewichteten Netzen nicht zur Verfügung. Ist dies ein Problem für den ersten

---

[12] Ein Überblick über die auf Wolfram Alpha verwendeten gewichteten Netzwerke ist abrufbar unter https://resources.wolframcloud.com/NeuralNetRepository. Die durch Lee verwendete Funktion zur Bilderkennung basiert auf einem eigens entwickelten gewichteten Netzwerk *Wolfram ImageIdentify Net V1*, über dessen Trainingsdaten der Hersteller auch auf Nachfrage keine Auskunft gibt.

[13] In Wordnet findet die Suche nach „art“ vier verschiedene Synsets (ID: 02746552, 00935235, 05646832, 07011408), von denen „artwork“ (ID: 07011408) für unsere Zwecke das relevanteste ist (vgl. http://wordnetweb.princeton.edu/perl/webwn?s=art&sub=Search+WordNet, abgerufen am 11.3.2021). Das Substantiv „painting“ (ID: 03882197, 00938436) ist ebenfalls mit vier Synsets vertreten, wobei sich zwei auf Malern als Handwerk beziehen, nicht auf Malerei (ID: 00718460, 00610504). Da einige Bilder laut Lee als Kunst erkannt wurden, wenn sie einen Rahmen besaßen, wurde auch „frame“ im Sinne von Bilderrahmen (ID: 03395829) gesucht.

[14] Zur Vorgehensweise: Für *ImageNet Fall 2011* gibt es eine Liste, in der die vorhandenen Synset-Kategorien per ID-Nummer aufgelistet sind (http://image-net.org/ImageNet_data/urls/ImageNet_fall11_urls.tgz, abgerufen am 13.3.2017). Diese sind jeweils mit Download-URLs verknüpft, von denen Trainingsdaten heruntergeladen werden können. Zu einem großen Teil betrifft das Bilder von Flickr, zum Teil unter Creative Commons Lizenz, teils auch unlizensiert. In dieser Liste waren die gesuchten IDs nicht auffindbar. ImageNet LSVRC 2012 beinhaltet 1000 Klassen mit 1,28 Millionen Bildern, darunter jedoch ebenfalls keine der gesuchten Klassen, siehe https://raw.githubusercontent.com/mf1024/ImageNet-Datasets-Downloader/master/classes_in_imagenet.csv (abgerufen am 11.3.2021). Gleiches gilt für die Sammlung ImageNet 21K mit 21.000 bebilderten Klassen, siehe https://github.com/dmlc/mxnet-model-gallery/blob/master/imagenet-21k-inception.md (abgerufen am 11.3.2021).

Prototypen von The Curator's Machine? Dieser verwendet derzeit keine Erkennung anhand der Labels, sondern vergleicht das Bildmaterial anhand von Ähnlichkeiten. Dies erfolgt anhand sogenannter Feature (mathematische Vektoren), welche bestimmte Muster innerhalb der Bilder detektieren.

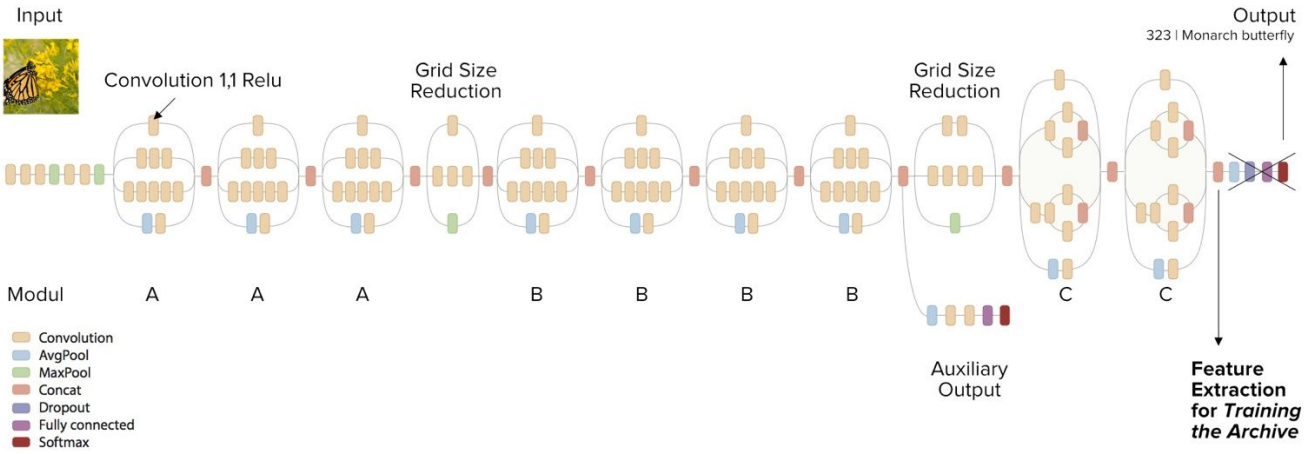


Abb. 3: Google InceptionV3 Architektur. Jedes abgerundete Rechteck stellt eine Convolution oder andere mathematische Funktion dar. Die Concats, welche die Module trennen, dienen der Reduzierung der Features. Siehe Farblegende. (der Verf. auf Basis von Google https://cloud.google.com/tpu/docs/tutorials/inception).

In Abb. 3 ist am Beispiel von InceptionV3 aufgezeigt, bis zu welchem Punkt das mit ImageNet trainierte Netz verwendet wird. Neben InceptionV3 setzt The Curator's Machine in ähnlicher Weise die per ImageNet trainierten Netze BiT/m-r152×4 und einzelne Module aus VGG19 in zwei Variationen ein.[15] InceptionV3 und BiT/m-r152×4 sind in The Curator's Machine mit der Bilddatensammlung ImageNet vortrainiert.

Wie sehen diese Features aus, lassen sie sich darstellen? Im Modul A1 gibt es in InceptionV3 beispielsweise eine Convolution (1,1 Relu, siehe Abb. 3), welche auf Vorhänge bzw. Faltenwurf anspricht. Da der Faltenwurf in der gotischen figürlichen Darstellung eine große Rolle spielte, soll dieser hier kurz verfolgt werden. Mit Hilfe des Werkzeugs Open AI Microscope[16] können Teile gewichteter Netze visualisiert werden. Dort wurde nach Faltenwurf-Abbildungen gesucht und eine bestimmte Convolution – Unit 45 – identifiziert, die entsprechende Muster aktiviert (Abb. 4, Abb. 5, Abb. 6).[17] Diese Muster entsprechen den mathematischen Features, welche mit Hilfe der ImageNet Bildsammlung trainiert wurden.

[15] An dieser Stelle wird es komplex, denn die ebenfalls per ImageNet trainierten VGG19 Module wurden mit einer geringeren Gewichtung (90%) eingeführt, und zwar nach dem Prinzip des Style-Transfers, das heißt, sie detektieren jene Texturen, die von Computerwissenschaftlern als „Style" bezeichnet werden. Diese Bildstile sind in sich problematisch, da sie nicht auf kunsthistorischer Expertise beruhen, sondern einem Populärverständnis von Kunst folgen, hier aber im Gewand von Wissenschaft auftreten (vgl. Gatys, Ecker und Bethge 2015, 6).

[16] Siehe https://microscope.openai.com (abgerufen am 11.3.2021).

[17] In den folgenden Schichten des Netzwerkes konnte ein ähnlich starker Verweis auf Faltenwurf nicht mehr gefunden werden.

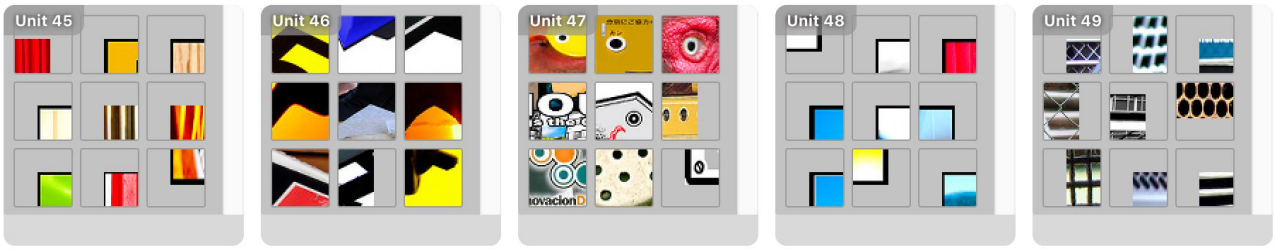


Abb. 4: Unit 45 (erste Kachel) der Convolution 1,1 Relu reagiert in InceptionV3 trainiert mit ImageNet u.a. auf Vorhänge. Die Darstellung zeigt Bildbeispiele aus dem ImageNet-Datenset (Open AI Microscope).

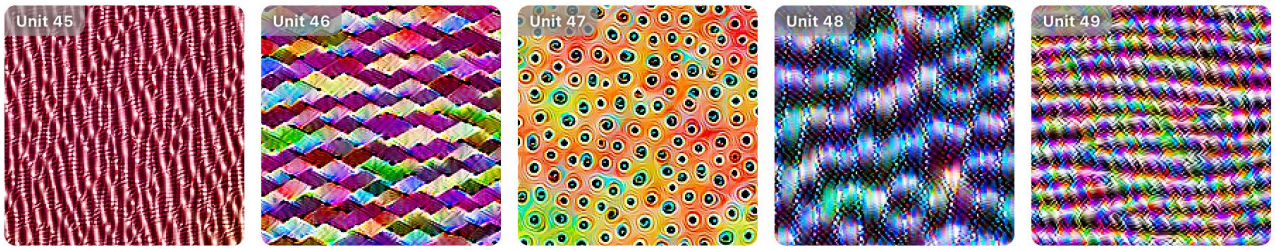


Abb. 5: Unit 45 (erste Kachel) – Diese Darstellung mit Hilfe des Deep-Dream-Algorithmus zeigt die Muster spezifischer Units in den Convolutionen der gewichteten Netze. Die erste Kachel, entspricht dem „Vorhang"-Muster (Open AI Microscope).

Für The Curator's Machine werden, wie oben beschrieben, die Features aus drei gewichteten Netzen zusammengeführt, bevor die Bilder der SMK Sammlung (Abb. 2) als Eingangsdaten darauf angewendet werden. Für jedes per ImageNet vortrainierte Netz werden die Bilder als Eingangsdaten eingegeben. Dann werden jeweils für jedes Bild die netzspezifischen Features extrahiert und diese dann am Ende zusammengeführt, um aus diesen den latenten Raum aufzubauen, wobei die Features vor dem Aufbau des Raumes noch einmal mit mathematischen Mitteln reduziert wurden.

Die Praktiker*innen der Computer Vision gehen derzeit davon aus, dass ImageNet trainierte Features genügend generalisieren, wenn in späteren Schichten noch eigens trainierte Features hinzugefügt werden (Yosinski u. a. 2014; Huh, Agrawal und Efros 2016, 6; Kornblith, Shlens und Le 2019). Huh u. a. stellten fest, dass die Abwesenheit von Klassen nur geringe Auswirkungen auf die Generalisierbarkeit von Features in ImageNet zeigen (ebd., 7). Hingegen zeigen Goh u. a. auf,[18] dass die ‚Neuronen' beziehungsweise Knoten der gewichteten Netze die Ontologie von ImageNet beziehungsweise WordNet nachbilden: „[…]es scheint, als ob sich die Neuronen in einer Taxonomie von Klassen anordnen, welche anscheindend die Imagenet-Hierarchie, zumindest annähernd, imitieren" (Goh u. a. 2021).

[18] Jenseits dieser Beobachtung ist der Text von Goh u. a. problematisch, denn er versucht, umstandslos bestimmte Gesichtszüge auf menschliche Emotionen abzubilden, und dies wiederum in den ‚Neuronen' gewichteter Netzwerke zu identifizieren. Gegen derartige phrenologische Illusionen sprechen sich unter anderem Bowyer u. a. in *The "Criminality from Face" Illusion* (Bowyer u. a. 2020), Stinson in *The Dark Past of Algorithms That Associate Appearance and Criminality* (Stinson 2020) und Munn in *Logic of Feeling – Technology's Quest to capitalize Emotion* (Munn 2020) aus.

Offert und Bell zeigen in ihrem Text *Perceptual bias and technical* auf, dass Daten-Bias auch in vortrainierten Netzen – also genau dem Anwendungsfall von The Curator's Machine – auftreten kann. Ihr Beispiel ist das Auftreten von Gittern („Fence"). Diese sollten eine Gitterzaunstruktur abbilden. Dafür analysierten sie 1300 Bilder aus der Fence-Kategorie in ImageNet und stellten fest, dass 1% bis 5% dieser Trainingsbilder Gefangene hinter Gittern zeigen und folgern: „Bilder von Menschen hinter Gittern erscheinen aus Sicht der Klassifizierung als besonders gitterähnlich" (Offert und Bell 2020, 9). Übertragen auf unser Beispiel kann an dieser Stelle nur vermutet werden, welche Art von Bias im Faltenwurf auftritt. Dies bedarf weiterer Untersuchung.

Bezogen auf den Prototypen von The Curtor's Machine ist daher anzunehmen, dass die Abwesenheit von Kunst keinen durchaus Einfluss auf die Generalisierbarkeit der Features hat.

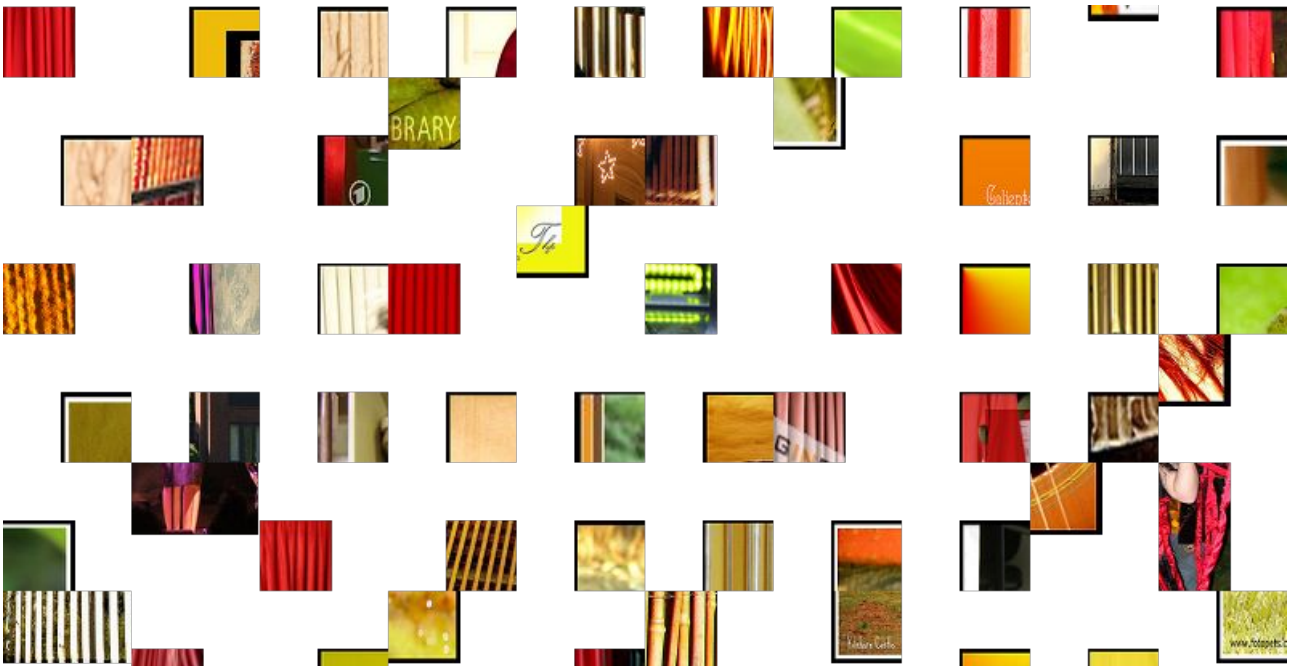


Abb. 6: Zoom auf Unit 45 (Ausschnitt): Bei genauerer Inspektion ist zu sehen, dass Vorhänge (rot), aber auch andere gestreifte Strukturen aktiviert werden, beispielsweise seriell angeordnete Zaunlatten oder Bambusrohre. Die Darstellung zeigt Bildbeispiele aus dem ImageNet-Datenset (Open AI Microscope).

Ist damit das Problem der Nicht-Klassifizierung von Kunst in ImageNet gelöst? Diese Frage ließe sich nur durch auf die Domäne der Kunst fokussierte Tests lösen. Grundsätzlich erlaubt es das Verfahren der ‚Backpropagation' beim Feature Transfer, die unteren Layer, zum Beispiel die ersten zwei A-Module (laut Abb. 3), anhand der oberen, transferierten Layer einem Fine-Tuning zu unterziehen, sodass die unteren Layer besser in Bezug auf die Domain ‚Kunst' generalisieren würden. Dieses Verfahren könnte zusätzlich zur Nutzung der unteren, vortrainierten Layer in Betracht gezogen werden (vgl. Yosinski 2014, 6).

Zwischenfazit: Kunst ist als Kategorie in ImageNet 2011 nicht annotiert und wenn, dann nur am Rande vorhanden. Computer-Vision-Projekte der Digital Humanities, die vortrainierte gewichtete Netze *zur Klassifikation* verwenden, sollten die Genealogien von ImageNet 2011 in ihre Überlegungen einbeziehen. Für Training the Archive ist zu überprüfen, ob die

unteren Layer vortrainierter Netze durch Backpropagation domain-spezifischer Layer verändert werden können. Die bisherige Forschung zu den Auswirkungen vortrainierter Netze, deren Features weiter verwendet werden, ist unabgeschlossen und kommt daher zu einander widersprechenden Aussagen. Für Training the Archive bedeutet dies, dass entsprechende Untersuchungen mit Bezug auf die eigenen Datensätze durchzuführen sind.

3.2 Fehlende Historizität

Auffallend in der Verwendung von ImageNet und auch anderen öffentlich verfügbaren Trainingsdaten wie Open Images Dataset und Microsoft COCO/Azure ist deren mangelhaftes ‚historisches Gedächtnis'. Die Eingangsdaten für den ersten Prototypen von The Curator's Machine, die Gemälde und Zeichnungen des Statens Museum for Kunst, stammen größtenteils aus Europa zwischen dem 15. Jahrhundert und dem 20. Jahrhundert.

Über diesen langen Zeitraum sind die inhaltlichen Änderungen der Bildsujets und der Darstellungsweisen signifikant. So unterscheidet sich die Körperdarstellung in der gotischen Malerei durch Betonung von Längen und Streckung der Gliedmaßen als Ausdruck höfischer Eleganz von heute vorherrschenden Körperbildern. Die Komplexität des Faltenwurfes hatte in der Gotik ihre ganz eigene Bedeutung, welche in heutigen Bilddaten selten anzutreffen ist.

Diese Punkte illustrieren spezifische Historizitäten der Eingangsdaten. Diese treffen nun auf die Zeitgenossenschaft der ImageNet-Trainingsdaten, welche für vortrainierte gewichtete Netze verwendet wurden. Die Trainingsdaten sind zu verschiedenen Zeitpunkten ab 2010 aus dem Internet geladen worden. Dazu wurden Suchmaschinen in fünf Sprachen anhand der Wordnet-Synsets nach Bildern durchsucht (Li u. a. 2009). Ein großer Anteil der Bilder stammt von der Fotografieplattform Flickr. Die Labels der ImageNet-Bilder wurden von prekären Clickworker*innen mittels der Plattform Amazon Mechanical Turk annotiert.

Objekte, die ab den 2010er Jahren fotografiert und klassifiziert wurden, trainieren nun also ein gewichtetes Netz, welches Eingangsdaten verarbeiten soll, die sich auf historische Formensprachen seit dem 15. Jahrhundert beziehen. Die zugrundeliegende Problematik wird auch von der Machine Learning Community selbst gesehen: „Diese Ergebnisse deuten darauf hin, dass Klassifikatoren, die auf modernen Techniken des Machine Learning basieren, [...], nicht die wahren zugrundeliegenden Konzepte lernen, welche ein korrektes Ausgabelabel erzeugen. Stattdessen haben diese Algorithmen ein Potemkinsches Dorf errichtet" (Goodfellow, Shlens und Szegedy 2015, 2).[19]

Diese Problematik wird in drei aktuellen künstlerischen Projekten sichtbar: *What the Machine Saw* (2019) von John Stack labelte eine Reihe von Abbildungen aus der Sammlung der Science Museum Group mit Hilfe des Amazon Rekognition Dienstes (Abb. 7). Diesen durch Rekognition automatisch vergebenen Labels, welche auf trainierten gewichteten Netzwerken beruhen, werden die Beschreibungen aus den Meta-Daten den Bildern gegenübergestellt. Die Metadaten wurden von Museumsmitarbeiter*innen bei Aufnahme der Gegenstände in die Sammlung erstellt (Stack 2019). Die Arbeit zeigt den Unterschied zwischen automatisierter

[19] Fortsetzung des Zitats: „This is particularly disappointing because a popular approach in computer vision is to use convolutional network features as a space where Euclidean distance approximates perceptual distance. This resemblance is clearly flawed if images that have an immeasurably small perceptual distance correspond to completely different classes in the network's representation" (ebd.).

Klassifikation, basierend auf den statistischen Modellen des Machine Learnings und den Annotationen, die durch Menschen als Meta-Daten erstellt wurden.

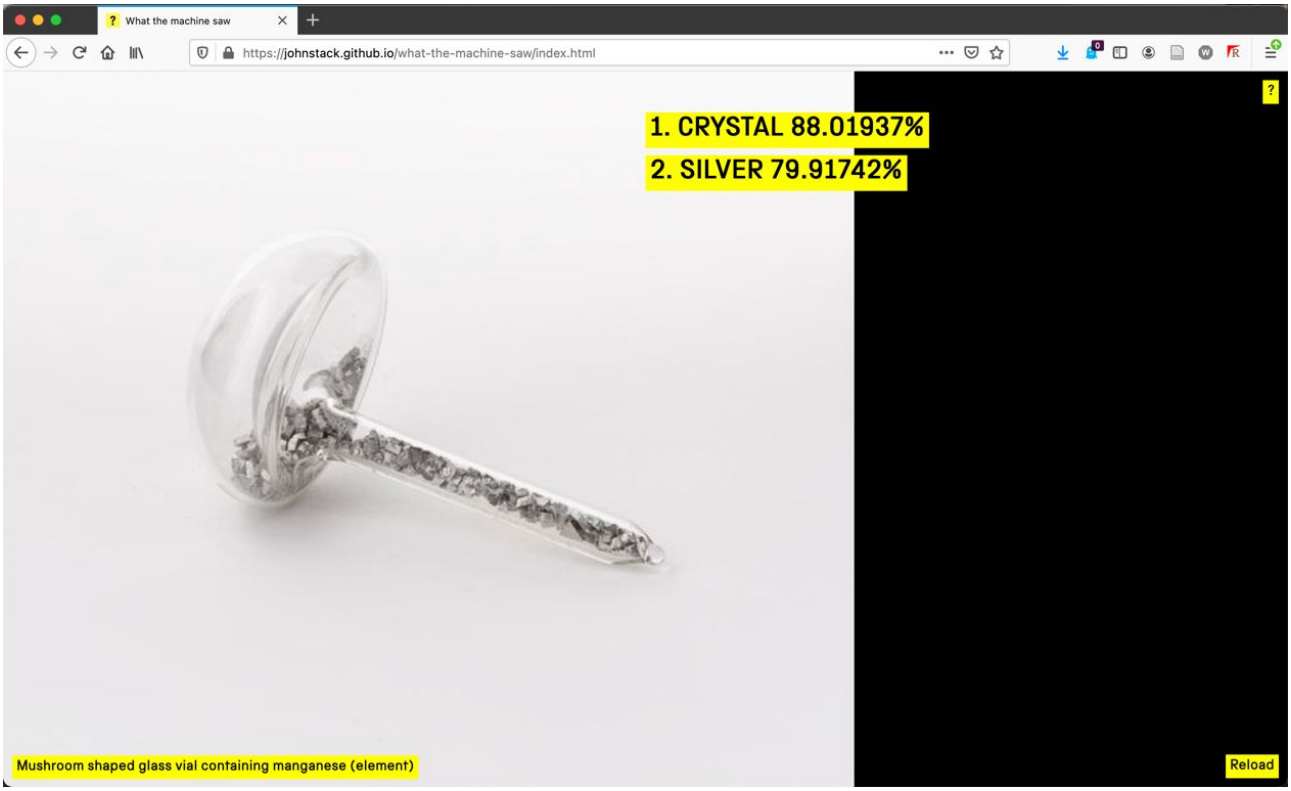


Abb. 7: John Stack *What the Machine Saw*: Ein pilzförmiges Glas, gefüllt mit Mangan wird als Kristall, beziehungsweise Silber erkannt (Screenshot).

Ein zweites künstlerisches Projekt stammt von Philipp Schmitt. *Declassifier* (2019–2020) überlagert eigens fotografierte Alltagsszenen aus New York mit den Bildern, welche für das Training des zugrundeliegenden Datensets Microsoft COCO verwendet wurden (Abb. 8). Im Beispiel ist eine Straßenszene aus Manhattan mit vorbeieilenden Passant*innen zu sehen. Wenn man mit der Maus über einen Objektrahmen fährt (violett), welcher ein durch Computer Vision erkanntes Objekt markiert, erscheint eines der Orginalfotos aus dem Trainingsdatensatz. Außerdem wird die Autorenschaft der Trainingsbilder, die im COCO-Datensatz ignoriert wird, erneut kenntlich gemacht, indem in einem weiß hinterlegten Informationskasten Autor*in, Titel und Dateinamen angegeben werden (Schmitt 2019). Schmitts Projekt demonstriert eindrucksvoll den Zusammenhang zwischen Trainingsdaten und Eingangsdaten und zeigt, wie verschiedenste räumliche, zeitliche und topologische Ordnungen aufeinanderprallen.

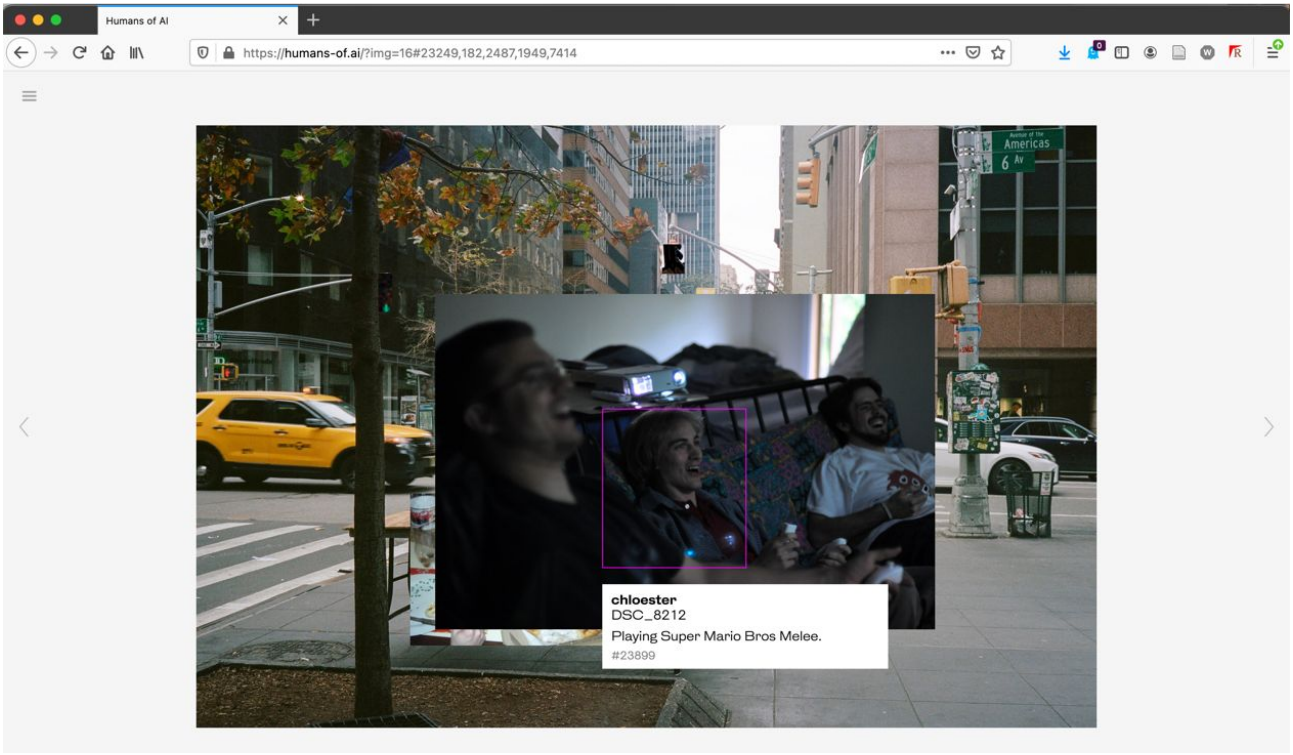


Abb. 8: Philipp Schmitt *Declassifier*: Das Foto einer Straßenszene in Manhattan, auf dem sich Personen befinden, wird überlagert mit Bildern, anhand derer die Objekterkennung „Person" traininert wurde, in diesem Falle ein Bild von „chloester" mit dem Titel „Playing Super Mario Bros Melee" (Screenshot).

In dem dritten Projekt, *Recoding Art*, wurde durch die Künstler Gabriel Pereira und Bruno Moreschi ein Teil der Sammlung des Van Abbemuseum Eindhoven mit 654 Abbildungen untersucht (Abb. 9). Alle Abbildungen aus der Sammlung sind freigestellt, ausgeleuchtet und farboptimiert, wodurch sie als operative Bilder gut geeignet sind. Es fehlen jedoch beispielsweise Größeninformationen und andere Metadaten. Mit Hilfe des selbst programmierten Werkzeugs *Recoding Art* untersuchten Pereira und Moreschi, wie Kunstwerke durch Computer Vision als Alltagsgegenstände interpretiert werden. *Recoding Art* gibt zu jedem einzelnen Bild die Interpretation durch eine Reihe von APIs aus: Google Cloud Vision, Microsoft Azure, Amazon Rekognition, IBM Watson, Facebook Detektron and Darknet YOLO (Pereira und Moreschi 2020, 2).

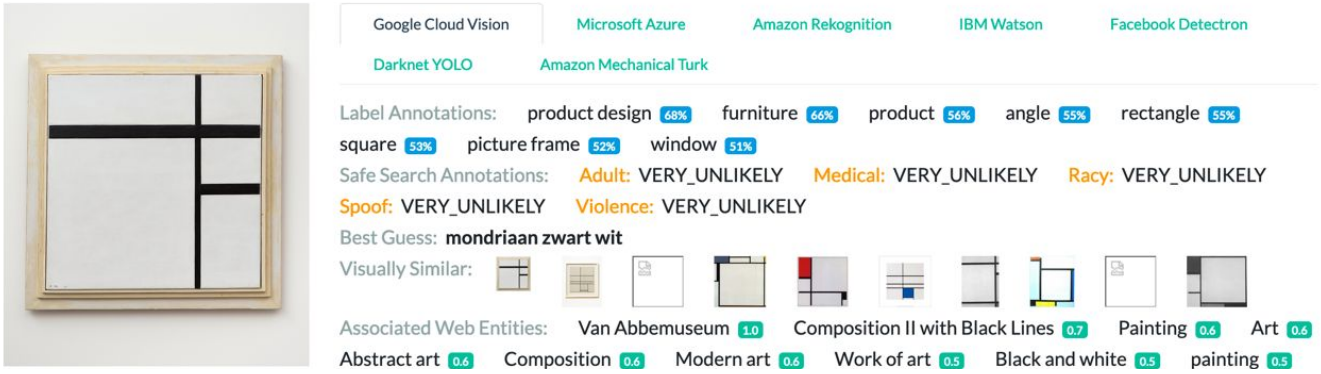


Abb. 9: Gabriel Pereira und Bruno Moreschi *Recoding Art*: Google Cloud Vision detektiert Piet Mondrians „Composition en blanc et noir II", 1930 (Screenshot, https://testingdataviz.github.io/VAD0.4/).

Google Cloud, Amazon Rekognition und IBM Watson sind vergleichsweise gut darin, ‚Kunst' oder ‚Gemälde' zu detektieren, vermutlich weil sie auch Meta-Daten und Suchergebnisse einbeziehen. Die anderen Dienste, detektieren einzelne Objekte („Person", „Tisch", „Baum") mit unterschiedlichem Erfolg, erkennen jedoch das Konzept „Kunst" als solches nicht oder nur mit niedriger Wahrscheinlichkeit.

Google Cloud Vision ist in der Lage, Skulpturen (z.B. Christos and Jeanne-Claudes *Wrapped Armchair* und Ernst Barlachs *Lehrender Christ*) als solche zu detektieren und nichtfigurative abstrakte Gemälde (Fernand Léger *L'accordéon*) als „Gemälde". Aus dieser Beobachtung ergibt sich die Frage, was Google Cloud Vision von den anderen APIs unterscheidet. Es ist zu vermuten, dass die Erkennung auf Werke zutrifft, die in Verbindung mit dem Google Arts and Culture Project stehen, welches zahlreiche Kunstwerke digitalisiert hat und diese in 13449 Künstler*innen, 240 Medien und 117 Kunstrichtungen (Stand 3.3.2021) kategorisiert hat.[20]

In Bezug auf die oben angesprochene Historizität konstatieren Moreschi und Pereira: „Die überwiegende Mehrheit der Werke (fast 90%) wurde in mindestens einem der Ergebnisse als Konsumprodukt gelesen, wie sie üblicherweise in Kaufhäusern zu finden sind" (ebd., 6). Dieser Befund ist nicht zu unterschätzen in Bezug auf das Verhältnis der zeitgenössischen Trainingsdaten zu den historischen Eingangsdaten der Bildsammlungen, die untersucht werden sollen. Die Autoren konstatieren, dass hinter den derzeit trainierten Netzen eine „kapitalistische Logik" (ebd., 12) am Werk sei, welche entsprechende Normativität reproduziere.[21]

Aus den hier aufgeführten drei Beispielen wird die Ahistorizität von Computer Vision als Problem erkennbar: Wenn in einem Gemälde des 15. Jahrhunderts ein Faltenwurf („drapery", 03237826) klassifiziert werden kann, dann weil der Faltenwurf mit einer operativen Abbildung aus dem (metaphorischen) „Ikea Produktkatalog" trainiert wurde. Dies lässt vermuten, dass eine ganze Reihe von Formen, Texturen, Objekten nicht trainiert werden, da sie in der heutigen Produktwelt nicht vorkommen, oder eben ahistorisch sind, denn ihre damalige Bedeutung entspricht nicht der heutigen Bedeutung.

In der Gotik war der ausgeprägte Faltenwurf stilistisch markant für Skulpturen und Malerei (Sauerländer 1970). Die Kunstgeschichte unterscheidet zwischen „Falten, Faltenkaskaden, Muldenfalten, Omegafalten, Parallelfalten, Röhrenfalten, Schüsselfalten, Tütenfalten, V-förmigen Falten, Y-förmigen Falten und Zackenstilfalten" (Kunsthistorisches Institut der CAU Kiel 2019). Derartige Taxonomien haben wenig mit jenen Falten der Gardinen und Vorhänge zu tun, die heute als warenförmige Produkte existieren.

Aus kunsthistorischer Perspektive muss hier eine deutliche Warnung vor falscher Objektivität der gewichteten Netze ausgesprochen werden.

Geht es jedoch um ein ‚Dekolonisieren' des kuratorischen Blickes, der nicht allein streng kunsthistorisch vorgeht, so können sich neue Assoziationen ergeben: Moreschi und Pereira schlagen vor, die Detektionen durch Computer Vision nicht allein als Problem wahrzunehmen. Stattdessen würde das Ignorieren von Autorenschaft und Historizität Potenziale eröffnen, neue Narrative der Kunstgeschichte durch neue Assoziationsketten, die über verschiedene geografische Regionen und zeitliche Perioden hinwegreichen, zu provozieren (Pereira und Moreschi 2020, 17). Dass die neuronalen Netze kunsthistorisch „valorisierte

---

[20] Siehe: https://artsandculture.google.com/explore (abgerufen am 15.3.2021).

[21] Piet Mondrians modernistische Komposition *Composition en blanc et noir II* von 1930 wird durch Google Cloud Vision beispielsweise als „product design, 68%; furniture 66%, picture frame 52%, window 51%" detektiert und On Kawaras *JULY 4, 1973 Wednesday* als „font 78%, product design 62%, brand 62%".

Werke“ (Groys 2004) unter Missachtung des Kanons ähnlichen Amateurkunstwerken gegenüberstellen, eröffne den Weg für die Wahrnehmung von Kunst jenseits des Museums (Pereira und Moreschi 2020, 20).

Zwischenfazit: Die ahistorisch und ageografisch[22] trainierten neuronalen Netzwerke ermöglichen ineinander verwobene, post-humanistische Mensch-Maschinefigurationen, die Raum für neue epistemische Prozesse schaffen. Ziel von Training The Archive ist es, diese Dekontextualisierung in einem ersten Schritt zuzulassen, um sie in einem zweiten Schritt einer erneuten menschlichen Bewertung zu unterziehen, und zwar als Kollaboration zwischen Kurator*in und Maschine. Beispielhaft wurde die Ahistorizität der Trainingsdaten im Verhältnis zu den Eingangsdaten aufgezeigt.

3.3 Textur und Umriss

Ein Ziel des ersten Prototypen von Training the Archive sind explorative visuelle Darstellungsformen wie Clusteranalyse, Gridplot oder Scatterplot für die Eingangsdaten. Diese können, wie im Abschnitt *Triplet-Loss-Function* von Bönisch ausgeführt, zusätzlich durch menschlichen Input trainiert werden (Bönisch 2021, 7–11).

Textur und Faktur stellen wichtige kunsthistorische und kuratorische Bewertungskriterien dar. Diese bestimmen allerdings nicht nur die repräsentative, sondern auch die operative Dimension von Bildern mit. Wie wirken sich diese auf die angestrebten Darstellungen in Cluster, Grid und Scatterplot aus? Textur und Faktur treten in ein Wechselspiel mit spezifischen Medialitäten historischer Bilder. Nicht nur sind die Medialitäten und Qualitäten der verwendeten Bildträger wie Papier oder Leinwand, Pigmente, Farben und Tuschen über die Jahrhunderte veränderlich, es treten auch Alterungsprozesse wie Vergilben, Abdunkeln, Ausbleichen, Flecken und Craquelé auf. Textur und Faktur haben je nach Alterungs- und Bearbeitungsprozess Folgen für die Rezeption.

Abb. 10: Wenn die Textur der Haut eines indischen Elefanten (a) auf eine Katze (b) übertragen wird, führt der Textur-Umriss-Konflikt dazu, dass die Katze als Elefant erkannt wird (c) (Geirhos u. a. S.1, Abb. 1).

[22] Ageographisch meint hier die unterschiedslose Vermischung verschiedenster visueller Kulturen in den Trainingssets, beispielhaft sichtbar im Open Images Dataset, welches asiatische und europäische Skulptur als Skulptur verhandelt und die kulturelle Genese der jeweiligen künstlerischen Techniken und Sujets ausblendet.

Geirhos et. al. haben kürzlich experimentell gezeigt, dass mit ImageNet konditionierte gewichtete Netzwerke ‚Textur' gegenüber ‚Umriss/Form' (engl.: shape) präferieren: „ImageNet-Objekterkennung könnte im Prinzip allein durch Texturerkennung erreicht werden" (Geirhos u. a. 2019, 2). Dieses Prinzip bezeichnen sie als Textur-Bias. Sie füllten beispielsweise den Umriss einer Katze mit der Textur eines Elefanten. Dieses Bild wurde von gewichteten Netzen zuverlässig als Elefant erkannt (Abb. 10).

Daraus schlussfolgern sie, dass es sich beim Textur-Bias um eine der für Computer Vision und Muster-Erkennung nicht unüblichen Abkürzungen handelt. Die Algorithmen-Daten-Systeme sind in Abkürzungen derart optimiert, dass sie ein von Menschen validiertes Ergebnis liefern, indem sie den kürzesten, das heißt mathematisch-ökonomisch optimalsten Lösungsweg einschlagen: „Wenn Textur ausreichend ist, warum sollte ein CNN [Convoluted Neural Network] dann noch viel mehr lernen?" (ebd., 9).[23]

Abb. 11: Wenn dieser Holzschnitt von Lucas Cranach d. J. von 1548, welcher Martin Luther darstellt, durch Computer Vision detektiert wird, können die Alterungsflecken oder die Struktur des Papierrandes durch Textur-Bias Zuordnungen erzeugen, die aus menschlicher Sicht unerwünscht sind (Ausschnitt, Statens Museum for Kunst, Kopenhagen, gemeinfrei, KKSgb5082).[24]

Wie ist mit dem Texture-Bias umzugehen? Erstens könnte die Textur von Craquelé, wäre denn ImageNet ausreichend dahingehend konditioniert, dazu dienen, ‚ältere' Malerei zuverlässiger zu erkennen. Zweitens, und mit größerer Dringlichkeit, wären die grundsätzlichen Implikationen zu überlegen. Geihhos u. a. zeigen, dass ein modifiziertes ImageNet, welches

[23] Siehe auch *Shortcut Learning in Deep Neural Networks* (Geirhos u. a. 2020).

[24] Die Abbildung dient der Illustration des Problems nach Geihos u. a. 2020. Ein Einzelnachweis für den Prototypen The Curator's Machine anhand der Abbildung wurde nicht geführt.

Umrisse/Form stärker betont, den Textur-Bias abmildert: „Wir zeigen auf, dass der Textur-Bias in Standard-CNNs überwunden und in Richtung eines Kontur-Bias verändert werden kann, wenn mit einem geeigneten Datensatz trainiert wird“ (ebd., 3).

Als Verfahren für eine Korrektur von Textur-Bias schlagen sie vor, mittels Style-Transfer eine spezifische ImageNet Bilddatensammlung als Trainingsdaten zu erzeugen.[25] Für den Style-Transfer wird mittels des Verfahren AdaIN (Huang und Belongie 2017) eine Reihe von Stilen auf die Trainingsbilder übertragen, die zu einer Hervorhebung von Konturen und Umrissen führen. Indem ein und dasselbe manipulierte Bild aus dem ImageNet-Datensatz in verschiedenen Stilen zum Training verwendet wurde, konnten die Autoren sicherstellen, dass nicht Textur, sondern Kontur/Form in die Gewichte des trainierten Netzwerkes eingeschrieben werden (Geirhos u. a. 2019, 5).[26] Warum dieser Aufwand? Aus der menschlichen Neurophysiologie ist bekannt, dass Menschen Bilder vorrangig anhand von Umriss/Kontur erkennen.

Abb. 12: Ölmalerei auf Holz von Lucas Cranach d. Ä.: Martin Luther von 1532, Ausschnitt mit Augenmerk auf Craquelé (Statens Museum for Kunst, Kopenhagen, gemeinfrei, KMSsp720).[27]

[25] Siehe https://github.com/rgeirhos/Stylized-ImageNet (abgerufen am 15.3.2021). Vortrainierte stilisierte gewichtete Netzwerke stellen Geirhos u. a. unter https://github.com/rgeirhos/texture-vs-shape zur Verfügung.
[26] Da sie das Verfahren für allgemeine Bildmengen vorschlagen, ist erwähnenswert, welche Datensammlung sie für den Styletransfer verwendeten: Kaggle's *Painter by Numbers*, welches größtenteils wiederum auf Malereiabbildungen, die auf WikiArt gesammelt wurden, basiert. Siehe https://www.kaggle.com/c/painter-by-numbers/ und https://www.wikiart.org (abgerufen am 15.3.2021).
[27] Wie vorhergehende Anmerkung. Zudem bleibt zu erforschen, inwiefern die feinteilige Craquelé bei Pixelgrößen von 244×244 Pixel überhaupt noch relevant ist.

Zwischenfazit: Der erste Prototyp von The Curator's Machine basiert auf mit ImageNet trainierten gewichteten Netzwerken. Ein Bias zugunsten von Textur und Faktur könnte für Zuordnungen sorgen, die für an Form/Umriss orientierte menschliche Betrachter*innen ungewöhnlich und neu sind, sodass hier auch andere Erkenntnisse möglich sind. Die Nutzer*innen sollten daher darauf hingewiesen werden, auf welche Weise Textur-Bias wirksam wird.

4 Fazit

Aus den Überlegungen zu den Trainingsdaten ist eine Reihe von Schlüssen zu ziehen:

1. Für die *Feature Extraction* ist zu testen, ob der nächste Prototyp mit gewichteten Netzen entwickelt werden kann, die nicht auf allein auf ImageNet basieren. Es ist zu überprüfen, in welchem Maße sich die Feature Extraction der vortrainierten Netze auf die darauf aufsetzenden Module von The Curator's Machine auswirkt. Andere Bilddatensammlungen, die in Bezug auf ‚Kunst' bessere Erkennungsleistungen aufweisen, zum Beispiel das durch Google initiierte *Open Images Dataset,* sollten auf ihre Eignung untersucht werden.
2. Die fehlende Historizität der vorhandenen Trainingsdatensätze ist eine semantische Schranke, welche, wenn überhaupt, nur durch Metadaten, die in die Netzwerke der Computer Vision eintrainiert werden, übersprungen werden kann. Hierfür kommt der Einsatz von Connecting Text and Images (CLIP) Netzwerken in Frage, die jedoch weitere Komplexität nach sich ziehen.
3. Klassische ImageNet trainierte Netze favorisieren Textur. Die Nutzer*innen von The Curator's Machine sind darauf aufmerksam zu machen. Idealerweise kann der Unterschied Textur versus Kontur als Option zu Auswahl gestellt werden.

Offene Forschungsfragen:

Mit Open Images und anderen, kleineren Datensets kann im Transferlearning getestet werden, in welchem Maße bestimmte kunsthistorisch relevante Features zweidimensionaler Werke erlernt werden können. Der Ahistorizität entledigt sich das gewichtete Netzwerk grundsätzlich nicht, da es eng mit der nicht-semantischen Gewichtung und den daraus abgeleiteten Vektoren vermengt ist. Mehrdimensionale Medien, auch zeitbasierte, ephemere oder konzeptuelle Strategien sind nur eingeschränkt durch Computer Vision zu verarbeiten. Hierfür sind entweder Umgangsweisen zu finden, oder die Ausschlüsse deutlicher zu markieren.

Als weitere offene Frage bleibt, in welchem Maße die hier ausgeführten Effekte der Ahistorizität von ImageNet in den Grids, Scatterplots und anderen visuellen Zuordnungsdiagrammatiken tatsächlich zum Ausdruck kommen. Dies wäre die Aufgabe einer spezialisierten Untersuchung.

Die ursprünglich repräsentativen Bilder, die in Datensets zu operativen Bildern werden, ändern ihren Status in den diagrammatischen Konstellationen der Rasterdarstellungen, Scatterplot, Nearest Neighbor und Pfaddarstellungen. Den Diagrammatiken wird sich daher ein späteres Working Paper widmen.[28]

[28] Der Autor dankt den folgenden Personen für ihre Kommentare: Inke Arns, Dominik Bönisch, Matthias Pitscher, Nicolas Malevé und Alexa Steinbrück.

LITERATURVERZEICHNIS

Alashhab, Samer, Antonio-Javier Gallego und Miguel Ángel Lozano. 2019. Hand Gesture Detection with Convolutional Neural Networks. In: Distributed Computing and Artificial Intelligence, 15th International Conference, hg. von Fernando De La Prieta, Sigeru Omatu, und Antonio Fernández-Caballero, 45–52. Cham: Springer International Publishing.

Amat, Josep und Alicia Casals. 1992. Image Obtention and Preprocessing. In: Computer Vision: Theory and Industrial Applications, hg. von Carme Torras, 1–58. Berlin, Heidelberg: Springer.

Bönisch, Dominik. 2021. The Curator's Machine. Clustering von musealen Sammlungsdaten durch Annotieren verdeckter Beziehungsmuster zwischen Kunstwerken. Training the Archive – Working Paper, Aachen/Dortmund. 10.5281/zenodo.4604880.

Bowker, Geoffrey C. und Susan Leigh Star. 1999. Sorting Things Out – Classification and Its Consequences. Cambridge, MA: MIT Press.

Bowyer, Kevin W., Michael C. King, Walter J. Scheirer und Kushal Vangara. 2020. The "Criminality From Face" Illusion. IEEE Transactions on Technology and Society 1, Nr. 4 (Dezember): 175–183. 10.1109/TTS.2020.3032321.

Broeckmann, Andreas. 2016. Machine art in the Twentieth Century. Leonardo Book Series. Cambridge, MA: MIT Press.

Brownlee, Jason. 2019. Best Practices for Preparing and Augmenting Image Data for CNNs. Blog. Machine Learning Mastery. 2. Mai. https://machinelearningmastery.com/best-practices-for-preparing-and-augmenting-image-data-for-convolutional-neural-networks (zugegriffen: 5. März 2021).

Burrell, Jenna. 2016. How the Machine 'thinks' – Understanding Opacity in Machine Learning Algorithms. Big Data & Society 3, Nr. 1 (5. Januar). 10.1177/2053951715622512.

Cardon, Dominique, Jean-Philippe Cointet und Antoine Mazieres. 2018. Neurons spike back: The Invention of inductive Machines and the Artificial Intelligence Controversy. Reseaux 36, Nr. 211: 173–220. 10.3917/res.211.0173.

Chaki, Jyotismita und Nilanjan Dey. 2020. A Beginner's Guide to Image Preprocessing Techniques. Boca Raton London: CRC PRESS.

Crawford, Kate und Trevor Paglen. 2019. Excavating AI – The Politics of Images in Machine Learning Training Sets. Website. Excavating AI. 19. September. https://www.excavating.ai (zugegriffen: 12. Mai 2020).

Gatys, Leon A., Alexander S. Ecker und Matthias Bethge. 2015. A Neural Algorithm of Artistic Style. arXiv (2. September). http://arxiv.org/abs/1508.06576.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann und Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture – increasing shape bias improves accuracy and robustness. arXiv (14. Januar). http://arxiv.org/abs/1811.12231.

Goh, Gabriel, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford und Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. Distill 6, Nr. 3 (4. März). 10.23915/distill.00030, https://distill.pub/2021/multimodal-neurons (zugegriffen: 24. April 2021).

Goodfellow, Ian J., Jonathon Shlens und Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv (20. März). http://arxiv.org/abs/1412.6572.

Graham, Beryl und Sarah Cook. 2010. Rethinking Curating – Art after New Media. Leonardo. Cambridge, MA: MIT Press.

Grau, Oliver, Janina Hoth und Eveline Wandl-Vogt. 2019. Digital Art through the looking Glass – New Strategies for Archiving, Collecting and Preserving in Digital Humanities. Krems a.d. Donau: Edition Donau-Universität.

Groys, Boris. 2004. Über das Neue – Versuch einer Kulturökonomie. 3. Aufl. Fischer Forum Wissenschaft Kultur & Medien. Frankfurt am Main: Fischer-Taschenbuch-Verl.

Hanna, Alex, Emily Denton, Razvan Amironesei, Andrew Smart und Hilary Nicole. 2020. Lines of Sight. Logic Magazine. Dezember. https://logicmag.io/commons/lines-of-sight (zugegriffen: 26. Februar 2021).

Hayles, Katherine. 2005. Computing the Human. Theory, Culture & Society 22, Nr. 1: 131–151. 10.1177/0263276405048438.

Huang, Xun und Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. arXiv (30. Juli). http://arxiv.org/abs/1703.06868.

Huh, Minyoung, Pulkit Agrawal und Alexei A. Efros. 2016. What makes ImageNet good for transfer learning? arXiv (10. Dezember). http://arxiv.org/abs/1608.08614.

Kornblith, Simon, Jonathon Shlens und Quoc V. Le. 2019. Do Better ImageNet Models Transfer Better? arXiv (17. Juni). http://arxiv.org/abs/1805.08974.

Krizhevsky, Alex, Ilya Sutskever und Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Volume 1, 1097–1105. NIPS'12. USA: Curran Associates Inc.

Kunsthistorisches Institut der CAU Kiel. 2019. Fachausdrückezur Benennung und Beschreibung von Gewandfalten figürlicher Darstellungen–Bildkünste. CAU Kiel. https://www.kunstgeschichte.uni-kiel.de/de/infos-fuer-das-studium/fachausdrucke-bildkunste-ss-2019-neu.pdf (zugegriffen: 3. März 2021).

Lee, Rosemary. 2020. Machine Learning and Notions of the Image. Dissertation, Copenhagen: Center for Computer Games Research, Department of Digital Design, IT-University of Copenhagen. https://en.itu.dk/~/media/en/research/phd-programme/phd-defences/2020/phd-thesis-final-version-rosemary-lee-pdf.pdf?la=en.

Li, Fei-Fei, Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, und Kai Li. 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. Miami, FL: IEEE, Juni. 10.1109/CVPR.2009.5206848, https://ieeexplore.ieee.org/document/5206848 (zugegriffen: 12. Mai 2020).

Malevé, Nicolas. 2020. On the data set's ruins. AI & SOCIETY. 10.1007/s00146-020-01093-w.

Munn, Luke. 2020. Logic of Feeling – Technology's Quest to capitalize Emotion. Lanham: Rowman & Littlefield.

Noble, Safiya Umoja. 2018. Algorithms of Oppression – How Search Engines reinforce Racism. New York: New York University Press.

Offert, Fabian und Peter Bell. 2020. Perceptual Bias and technical Metapictures – Critical Machine Vision as a Humanities Challenge. AI & SOCIETY (12. Oktober). 10.1007/s00146-020-01058-z.

Parikka, Jussi. 2021. On Seeing Where There's Nothing to See – Practices of Light beyond Photography. In: Photography off the Scale – Technologies and Theories of the Mass Image, hg. von Jussi Parikka, 185–210. Edinburgh: Edinburgh University Press.

Pasquinelli, Matteo. 2019. How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence. Spheres. Journal for Digital Cultures., Nr. 5 (November): 1–17.

Pereira, Gabriel und Bruno Moreschi. 2020. Artificial Intelligence and Institutional Critique – Unexpected Ways of seeing with Computer Vision. AI & SOCIETY (14. September). 10.1007/s00146-020-01059-y.

Rosenblatt, Frank. 1957. The Perceptron – A perceiving and recognizing Automation. Buffalo, NY: Cornell Aeronautical Laboratory.

Sauerländer, Willibald. 1970. Gothic Sculpture in France, 1140-1270. New York, NY: Harry N. Abrams.

Schmitt, Philipp. 2019. Declassifier. Website. Humans of AI. https://humans-of.ai (zugegriffen: 9. März 2021).

Stack, John. 2019. What the Machine saw. Website. https://johnstack.github.io/what-the-machine-saw (zugegriffen: 9. März 2021).

Stinson, Catherine. 2020. The Dark Past of Algorithms that associate Appearance and Criminality – Machine Learning that links Personality and physical Traits warrants critical review. Online Publication. American Scientist. 2. Dezember. https://www.americanscientist.org/article/the-dark-past-of-algorithms-that-associate-appearance-and-criminality (zugegriffen: 24. April 2021).

Yosinski, Jason, Jeff Clune, Yoshua Bengio und Hod Lipson. 2014. How transferable are Features in Deep Neural Networks? arXiv (6. November). http://arxiv.org/abs/1411.1792.