Deep Learning zur Analyse von Bildern von Seidenstoffen für Anwendungen im Kontext der Bewahrung des kulturellen Erbes

MAREIKE DOROZYNSKI[1], DENNIS WITTICH[1] & FRANZ ROTTENSTEINER[1]

Zusammenfassung: Der vorliegende Beitrag befasst sich mit der Klassifikation von Bildern mittels Deep Learning. Exemplarisch wird hier die bildbasierte Prädiktion der Entstehungsepoche von Seidenstoffen und Fertigungsskizzen betrachtet. An diesem Beispiel wird die Problemstellung von nicht eindeutig definierten Klassen sowie uneindeutigen Trainingsbeispielen aufgegriffen und ein Lösungsansatz vorgestellt. Dieser basiert auf einer erweiterten Verlustfunktion, mit welcher die uneindeutigen Trainingsbeispiele während des Trainings genutzt werden können. Hierzu werden während der Trainingsphase übergreifende Klassen definiert. Die durchgeführten Experimente zeigen, dass hiermit die Genauigkeit der Klassifikation ohne die übergreifenden Klassen verbessert werden kann. Da der exemplarische Datensatz relativ wenig Daten umfasst wird ein vortrainiertes Convolutional Neural Network verwendet und eine Augmentierung der Daten durchgeführt. Der Einfluss der Augmentierung sowie der entwickelten Verlustfunktion wird in einer Kreuzvalidierung evaluiert.

1 Einleitung

Für die Bewahrung des kulturellen Erbes sind Kunstsammlungen von hoher Bedeutung. Ebenso wichtig wie die eigentlichen Objekte sind hierbei die zugehörigen Informationen, wie beispielsweise das Datum der Entstehung, der Name des Künstlers oder die Stilrichtung. In aller Regel ist die Klassifikation von Kunstobjekten (im Sinne der Generierung oder Vervollständigung der zugehörigen Informationen) eine Aufgabe, die bislang nur von Fachleuten wie Kunsthistorikern bewältigt werden kann. Folglich erweist sich die Klassifikation mehrerer oder gar aller Objekte einer Sammlung oft als eine langwierige Aufgabe. Hieraus ergibt sich der Wunsch, die Klassifikation weitestgehend zu automatisieren. Bei ausreichender Anzahl klassifizierter Objekte ist es naheliegend, Verfahren des maschinellen Lernens heranzuziehen. Liegen beispielsweise digitale Bilder der Kunstwerke vor und ist der Hersteller von ausreichend vielen Werken bekannt, so kann ein Modell auf diesen Daten trainiert werden, das nach Abschluss des Trainings im Idealfall auch den Hersteller von weiteren Werken anhand von Bildern prädizieren kann.
Mit dieser Motivation hat sich das EU-Projekt SILKNOW unter anderem das Ziel gesetzt, eine automatische Methode für die bildbasierte Klassifikation von Seidenstoffen zu entwickeln. Die zu prädizierenden Variablen sind hierbei beispielsweise der Hersteller, die Stilrichtung oder die Herstellungsepoche des jeweiligen Seidenstoffes. In der vorliegenden Arbeit, die im Rahmen des Projektes entwickelt wurde, wird eine Methode für die bildbasierte Klassifikation vorgestellt. Als Datensatz dient exemplarisch eine Sammlung von Bildern von Seidenstoffen sowie zugehöriger Skizzen die für das Projekt zur Verfügung gestellt wurden (siehe Abbildung 1). Für alle Objekte

[1] Leibniz Universität Hannover, Institut für Photogrammetrie und GeoInformation, Nienburger Str. 1, D-30167 Hannover, E-Mail: [dorozynski, wittich, rottensteiner]@ipi.uni-hannover.de

der Sammlung liegen neben den Bildern auch gesammelte Informationen vor, z.B. über die Herstellungsepoche, also die zeitliche Einordnung der Seidenstoffe in Bezug auf die Fertigstellung. Die Herstellungsepoche wird in diesem Beitrag als zu prädizierende Variable gewählt, da sich hieraus eine Problematik ergibt, die über die ‚klassische' Klassifikation von Bildern hinausgeht. In der ‚klassischen' Klassifikation werden vorab disjunkte, eindeutige Klassen definiert, was jedoch bei der Prädiktion von Epochen, im Sinne von zeitlichen Abschnitten, aufgrund der Heterogenität der vorliegenden Annotationen oft nicht möglich ist. Auch in dem hier betrachteten Datensatz, in welchem die Angaben zu den Entstehungsepochen von Jahresangaben bis hin zu einer Zeitspanne von über 500 Jahren reichen und sich teilweise überlappen, ist die Definition von diskunkten Klassen nicht möglich. Hieraus ergibt sich das Problem von uneindeutigen Trainingsbeispielen bzw. Klassenlabels der Bilder. Hierunter verstehen sich Trainingsbeispiele, deren gegebene Annotation der Epoche nicht vollständig in einen Zeitraum fällt, der als Klasse definiert wurde.
Eine weitere Problematik, die in diesem Beitrag adressiert wird, ist das Training auf einem relativ kleinen Datensatz. Dies ist insbesondere bei tiefen neuronalen Netzen ein Problem, da diese oft mehrere Millionen Parameter aufweisen, deren Optimierung eine ausreichende Menge an Trainingsbeispielen erfordert.
Der hier gewählte Lösungsansatz basiert auf einem Convolutional Neural Network (CNN), das teilweise auf dem ImageNet-Datensatz (DENG et al. 2009) vortrainiert wurde. Die Verwendung eines vortrainierten Netzwerkes zur Merkmalsextraktion ist ein Ansatz, der sich in der Vergangenheit insbesondere für das anschließende Training auf kleinen Datensätze etabliert hat. Für den Umgang mit uneindeutigen Trainingsbeispielen wird eine erweiterte Verlustfunktion entwickelt, sodass auch diese für das Training des Netzwerkes genutzt werden können.
In Abschnitt 2 wird zunächst erläutert, welche Lösungsansätze für die bildbasierte Kunstklassifikation existieren und es wird ein Überblick über den bisherigen Umgang mit uneindeutigen Trainingsbeispielen gegeben. Anschließend wird in Abschnitt 3 der verwendete Datensatz vorgestellt. In Abschnitt 4 werden die Grundlagen beschrieben, auf welchen die Methodik in Abschnitt 5 aufbaut. In Abschnitt 6 wird der Ansatz experimentell evaluiert. Zuletzt wird der vorliegende Beitrag in Abschnitt 7 zusammengefasst und die wesentlichen Folgerungen werden herausgestellt.

2 Stand der Forschung

Die automatische, bildbasierte Klassifikation, Bewertung und Beurteilung von Kunstwerken ist eine Problemstellung, die bereits in zahlreichen Publikationen adressiert wurde. Frühere Arbeiten wie (BLESSING & WEN 2010) basieren oft auf manuell selektierten Merkmalen wie Histogramme der Farbwerte und Gradienten, die anschließend mit klassischen Klassifikationsmodellen, hier eine Support Vector Machine, klassifiziert werden. Die Autoren erreichen mit diesem Ansatz eine Genauigkeit von 85,1 % bei der Prädiktion des Künstlers.
Im Zuge der Weiterentwicklung der künstlichen neuronalen Netze (ANNs) und insbesondere der CNNs wurden auch diese Modelle auf kunstbezogene Klassifikationsprobleme angewendet. Beispielsweise verwenden AYUB et al. (2009) ein CNN zur Prädiktion des Preises von Gemälden. Sie erreichen hierbei jedoch nur geringe Genauigkeiten, was sie selbst unter anderem damit be-

gründen, dass der verwendete Datensatz mit 700 Bildern zu wenig Trainingsbeispiele beinhaltet. SUR & BLAINE (2017) zeigen jedoch, dass die Klassifikation von Kunstwerken mit CNNs prinzipiell möglich ist. Sie erreichen bei der Prädiktion der Künstler (25 verschiedene) unterschiedlicher Gemälde Genauigkeiten über 80 %. Im Gegensatz zu AYUB et al. Trainieren sie ihr Modell auf einem größeren Datensatz mit 13574 Bildern. Zudem nutzen sie zur Extraktion der Merkmale ein CNN, welches zuvor auf dem ImageNet Datensatz trainiert wurde. RAZAVIAN et al. (2014) zeigen, dass der Ansatz ein vortrainiertes Netzwerk für die Merkmalsextraktion zu verwenden, insbesondere für das Training auf kleinen Datensätzen sinnvoll ist. Ein verwandter Ansatz ist die Feinanpassung der Parameter eines bestehenden Modelles (Finetuning). Hierbei werden die Parameter eines vortrainierten Netzwerkes zur Initialisierung verwendet und während des Trainings auf dem Zieldatensatz weiter optimiert. Diesen Ansatz verfolgen auch HICSONMEZ et al. (2017), die ein Modell für die Prädiktion des Illustrators von Kinderbüchern trainieren. Da der dort verwendete Datensatz nur 6468 Bilder umfasst, verwenden die Autoren eine Datenaugmentierung, mit deren Hilfe sie eine Klassifikationsgenauigkeit von 90,0 % erreichen. Neben der Prädiktion des Künstlers befassen sich einige Arbeiten mit der Klassifikation von Kunstwerken nach Stil bzw. Genre (BAR et al. 2014; TAN et al. 2016; HENTSCHEL et al. 2016). In diesen Ansätzen wird verdeutlicht, dass die Verwendung von vortrainierten Netzwerken stets geringere Klassifikationsfehler erzielt als das Trainieren eines zufällig initialisierten Netzes.
In der Recherche zu dieser Arbeit konnte die Klassifikation von Seidenstoffen in keiner Arbeit ausgemacht werden. Lediglich XIAO et al. (2018) befassen sich mit der automatischen Klassifikation von Strickmustern mit künstlichen neuronalen Netzen. Das Modell ist hierbei auf hochauflösenden Nahaufnahmen der Gewebe trainiert, wodurch sich das Problem deutlich von dem hier behandelten unterscheidet.
Obgleich sich alle genannten Arbeiten mit der bildbasierten Klassifikation von Kunst befassen, verbleiben zwei signifikante Unterschiede gegenüber der vorliegenden Arbeit. Der erste Unterschied liegt darin, dass die Datensätze, die in den vorgestellten Arbeiten herangezogen wurden, überwiegend Digitalisierungen umfassen, die durch Scanverfahren oder in Form von rektifizierten Fotografien mit frontaler Beleuchtung erzeugt wurden. In dieser Arbeit wird der Umgang mit Fotografien der Kunstwerke behandelt, die starke Schwankungen sowohl in der Beleuchtung als auch in der Platzierung des eigentlichen Objektes innerhalb der Abbildung aufweisen. Der zweite Unterschied liegt in der Wahl der zu prädizierenden Klassen. Keine der Arbeit befasst sich mit dem zuvor erläuterten Problem der uneindeutigen Klassenlabels. Hierfür existieren einige Ansätze, in denen zusätzliche, übergreifende Klassen definiert werden und somit eine Art hierarchische Klassenstruktur gebildet wird. Einen Ansatz, basierend auf Support Vector Machines, stellen VURAL et al. (2004) vor. Einen ANN basierten Ansatz liefern WU et al. (2017). Beide Ansätze unterscheiden sich von dem hier vorgestellten Ansatz, da in diesem Beitrag die übergreifenden Klassen nicht prädiziert werden sollen. Stattdessen sollen die uneindeutigen Trainingsbeispiele für die Verbesserung der Prädiktion der Basisklassen genutzt werden.

3 Datensatz

Der hier genutzte Datensatz wurde von GARÍN zur Verfügung gestellt. Er enthält 2752 Bilder von Seidenstoffen und 2168 Bilder von Entwurfsskizzen (siehe Abbildung 1). In beiden Fällen ist

meistens sowohl eine Abbildung der Vorderseite als auch der Rückseite sowie gelegentlich auch eine Detailaufnahme vorhanden. Die Bilder haben eine Größe zwischen 500 x 500 Pixeln und 4000 x 3500 Pixeln. Zudem existieren Datenblätter mit detaillierten Beschreibungen der abgebildeten Stoffe. Diese enthaltenen z.B. Angaben bezüglich des Materials, der Herstellungsepoche oder der Weberei. Letztere variiert jedoch nicht, da alle Objekte der Sammlung von der Weberei der Familie Garín stammen. Der Familienbetrieb produziert jedoch seit dem Jahr 1820, sodass Stoffe aus fast zwei Jahrhunderten vorliegen; das älteste abgebildete Objekt ist eine Skizze aus dem Jahr 1827 und das jüngste Werk stammt aus dem Jahr 2015.

Abb. 1: Beispiele für das Bild einer Vorderseite (links) und einer zugehörigen Rückseite (Mitte) eines Seidenstoffs sowie ein Beispiel für ein Bild einer Entwurfsskizze (rechts)

Um die Bilder im Rahmen einer Klassifikation der Epoche nutzen zu können, müssen die gegebenen zeitlichen Angaben zunächst einzelnen Klassen zugeordnet werden. Eine Herausforderung bei der Einordnung der Beschreibungen der Herstellungsepoche in eine Klassenstruktur ist die stark variierende Genauigkeit der Zeitangaben die sich teilweise überlappen (siehe Abbildung 2). Für manche Stoffe ist der genaue Tag der Herstellung bekannt, wohingegen für andere die Entstehungszeit lediglich auf ein Jahrhundert eingegrenzt werden kann. Überschneidungen gibt es beispielsweise bei den Zeiträumen “2. Hälfte 19. Jahrhundert” und “1889-1912”. Bei einer Klasseneinteilung in einzelne Jahre oder Jahrzehnte wären beim vorliegenden Datensatz nicht ausreichend viele Beispiele pro Klasse vorhanden, wie Abbildung 2 verdeutlicht.

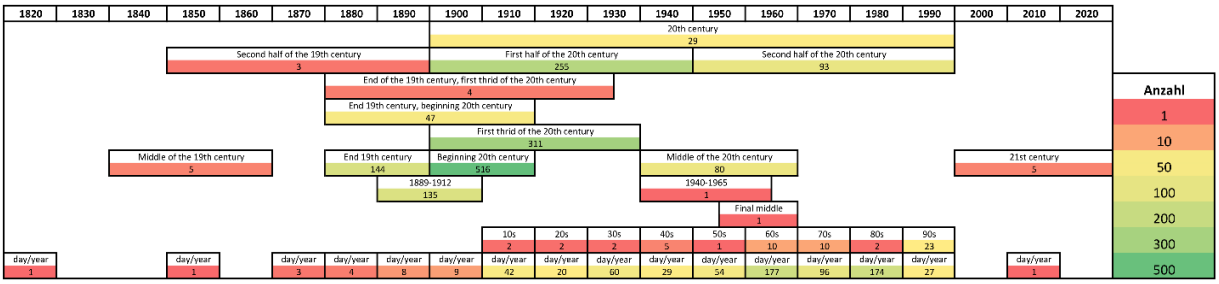

Abb. 2: Darstellung der vorliegenden Zeitangaben anhand eines Zeitstrahls, wobei der farblichen Codierung die Häufigkeit der jeweiligen Beschreibung zu entnehmen ist

Als Klassen werden daher im vorliegenden Beitrag halbe Jahrhunderte gewählt, sodass sich in Summe die drei Klassen “2. Hälfte 19. Jahrhundert”, “1. Hälfte 20. Jahrhundert” und “2. Hälfte 20. Jahrhundert” ergeben. Alle Bilder, deren Epochenangabe vollständig in einen dieser Zeiträume fällt, werden der entsprechenden Klasse zugeordnet. Die resultierende Klassenverteilung

ist in Tabelle 1 unter der Bezeichnung „Basisklassen“ dargestellt. Insgesamt lassen sich auf diese Weise 2876 der 4920 Bilder (ca. 58 %) einer der gewählten Klassen zuweisen, wobei sich diese Menge aus 2012 Seidenstoffbildern und 864 Skizzenbildern zusammensetzt. Zusätzlich werden Seidenstoffbilder, die in Bezug auf die drei Basisklassen eine uneindeutige Herstellungszeit haben sich jedoch zwei der Basisklassen umfassenden Zeiträumen von 100 Jahren zuordnen lassen, entsprechenden zusammengesetzten Klassen zugewiesen (siehe Tabelle 1).

Tab. 1: Übersicht über die Verteilung der Bilder auf die Klassen

	Basisklassen			**Zusammengesetzte Klassen**	
Klassenbezeichnung	2. Hälfte 19. Jh.	1. Hälfte 20. Jh.	2. Hälfte 20. Jh.	1850 -1950	1900 -2000
Anzahl Bilder	213	2035	628	302	140

4 Grundlagen

4.1 Convolutional Neural Networks

Durch CNNs können sowohl Merkmale aus Bildern als auch die Abbildung dieser Merkmale auf eine Menge von Klassen gelernt werden (KRIZHEVSKY et al. 2012). Hierfür werden sogenannte Convolutional Layer verwendet, welche Faltungen auf den Bildern ausführen, wobei die Gewichte der Faltungsmatrizen im Training ermittelt werden. Da auf diese Weise nur lineare Abbildungen realisiert werden, werden nichtlineare Aktivierungsfunktionen in CNNs integriert. Ein Beispiel für solch eine Aktivierungsfunktion ist die ReLU-Aktivierung (Rectified Linear Unit), die negative Werte auf null abbildet und positive Werte unverändert lässt (BISHOP 2006).
Da tiefere Netze potentiell komplexere Merkmale lernen wird angenommen, dass tiefere Netze bessere oder zumindest genauso gute Ergebnisse liefern wie flache Netze (HE et al. 2016a). Um das Training solcher tiefen Netzwerke zu ermöglichen führen HE et al. (2016a) sogenannte ‚Residual Blocks‘ ein und optimieren diese durch pre-activations (HE et al. 2016b). Der schematische Aufbau eines Residual Blocks mit pre-activations ist links in Abbildung 3 zu sehen. Die Abbildungsfunktion R, welche durch einen Residual Block gelernt wird, ist die Summe des Inputs Xt und der Residualfunktion des Inputs. Dabei geht Xt direkt in Abbildungsfunktion über eine sogenannte ‚shortcut connection‘ ein und die Residualfunktion wird mittels Convolutional Layern modelliert. Vor jedem Gewichtslayer wird eine pre-activation durch eine Batch Normalization (BN) mit anschließender ReLU-Aktivierung realisiert. Durch eine derartige Modellierung der Abbildungsfunktion müssen im Training lediglich die Gewichte der Residualfunktion gelernt werden, was zu besseren Ergebnissen im Vergleich zu genauso tiefen trainierten Netzen ohne Residuallernen führt (HE et al. 2016b).

4.2 Training

Allgemein wird bei dem Training eines CNNs eine Verlustfunktion minimiert. Basierend auf dem Verlustwert werden durch ‚Backpropagation‘ die Gradienten der Parameter des Netzes ermittelt. Entsprechend der Gradienten werden die Gewichte iterativ aktualisiert um den Verlustwert zu minimieren. Werden nicht alle Trainingsdaten zugleich, sondern eine zufällige Teilmenge der Daten (Mini-Batch) hierfür genutzt, nennt man dies den stochastischen

Gradientenabstieg (BISHOP 2006). Im Rahmen dieses Beitrags wird die Softmax Kreuzentropie als Verlustfunktion verwendet. Für die Berechnung des Funktionswertes werden zunächst alle Softmax-Aktivierungen der Klassifikationsschicht ermittelt. Bei insgesamt K zu prädizierenden Klassen gibt es K Netzwerkausgaben. Die Softmax-Aktivierung y_k der k-ten Klasse für die Netzwerkeingabe x berechnet sich zu

$$y_k(x, w) = \frac{\exp\left(w_k^T \Phi\right)}{\sum_{j=1}^{K} \exp\left(w_j^T \Phi\right)}$$

Gl. 1

aus den Gewichten w_k und w_j des k-ten beziehungsweise j-ten Ausgabeknotens und den Aktivierungen Φ der vorherigen Schicht im Netzwerk (BISHOP 2006). Dabei kann y_k als die Wahrscheinlichkeit interpretiert werden, dass das Sample x zu der Klasse k gehört. Anschließend berechnet sich der Verlustwert $E(w)$ für alle n betrachteten Samples aus der Kreuzentropie der Softmax-Aktivierungen

$$E(w) = -\sum_{n}\sum_{k} t_{nk} \ln\left(y_k(\mathbf{x_n}, \mathbf{w})\right)$$

Gl. 2

wobei $t_{nk} \in \{0, 1\}$ eine Indikatorvariable ist, welche angibt, ob das n-te Sample zur k-ten Klasse gehört (BISHOP 2006).

5 Methodik

Basierend auf den in Abschnitt 4 beschriebenen Grundlagen wird in diesem Abschnitt die Methodik erläutert, welche für die Prädiktion der Entstehungsepoche von Seidenstoffbildern angewendet wird. Diese umfasst die verwendete Netzwerkarchitektur sowie die Vorprozessierung der Daten und die Erweiterung der Softmax Kreuzentropie.

5.1 Netzwerkarchitektur

Das in diesem Beitrag genutzte CNN ist in Abbildung 3 dargestellt. Für die Merkmalsextraktion wird ein Residual Network mit 152 Convolutional Layern (HE et al. 2016b) genutzt. Als Eingabe wird ein Bild mit einer Größe von 224 x 224 Pixeln und drei Kanälen benötigt. Das Netzwerk setzt sich aus 50 Residual Blocks zusammen (siehe Abb. 3). Die Ausgabe der Merkmalsextraktion ist ein 2048-dimensionaler Vektor. Als Klassifikationsteil des Netzes werden 3 voll vernetzte Schichten mit ReLU- bzw. Softmax-Aktivierung verwendet. Für die Prädiktion der Epochen besteht der letzte Layer aus drei Knoten. Auf diese Weise werden die 2048 Merkmale auf die Menge der drei Klassen abgebildet.

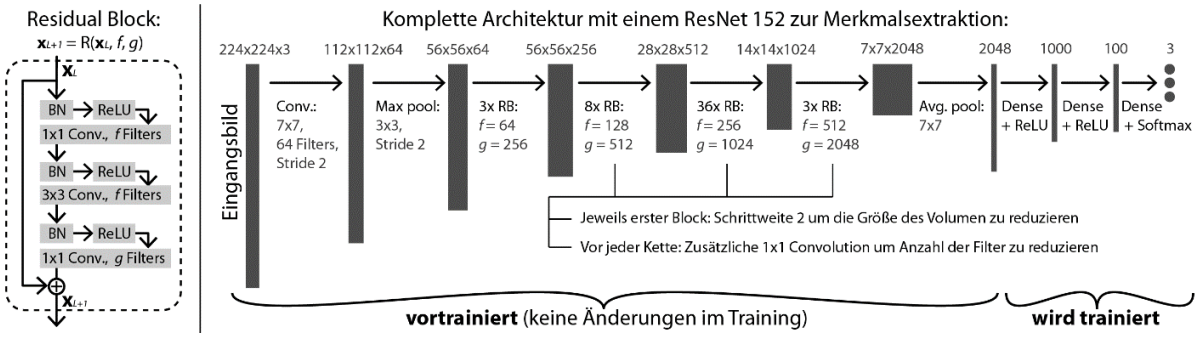

Abb. 3: Schematische Darstellung eines Residual Blocks (links) sowie die vollständige Netzwerkarchitektur (rechts).

5.2 Training

Für die im vorangegangenen Abschnitt vorgestellte Architektur wird aufgrund der geringen Menge an Trainingsdaten der Merkmalsextraktionsteil mit den auf dem ILSVRC-2012-CLS ImageNet-Datensatz (RUSSAKOVSKY et al. 2015) trainierten Gewichten des ResNet-152 V2 initialisiert. Diese Gewichte bleiben während des Trainings unverändert und lediglich die Gewichte des Klassifikationsteils werden auf dem Seidenstoffdatensatz trainiert. Bei der Nutzung üblicher Verlustfunktionen im Training muss hierfür eine eindeutige Zuordnung aller Samples zu einer der definierten Klassen gegeben sein, damit sie im Training verwendet werden können. Damit im Rahmen des Trainings auch Samples mit uneindeutigen Klassenlabels verwendet werden können, wird die Verlustfunktion aus Gleichung 2 erweitert. Dafür sei zunächst eine hierarchische Klassenstruktur mit zwei Ebenen von Klassen wie links in Abbildung 4 gegeben. Die untere Ebene besteht aus K Basisklassen $\{C_1, \dots, C_k, \dots, C_K\}$, welche durch den zu trainierenden Klassifikator prädiziert werden können. Sei hierfür $B \coloneqq \{1, \dots, k, \dots, K\}$ die Menge der Indizes der Basisklassen. Zudem gibt es eine zweite Ebene von L zusammengesetzten Klassen $\{C_{Q1}, \dots, C_{Ql}, \dots, C_{QL}\}$, wobei sich jede dieser Klassen aus der Vereinigung einer Teilmenge der Basisklassen zusammensetzt. Die Verlustfunktion $E(w)$ aus Gleichung 2 kann derart erweitert werden, dass die N im Mini-Batch enthaltenen Samples einer Basisklasse k oder einer aus den Basisklassen q zusammengesetzten Klasse Q_l zugehörig sein können. Die in die Kreuzentropie eingehenden Softmax-Aktivierungen der Basisklassen y_k berechnen sich weiterhin gemäß Gleichung 1, wohingegen die Softmax-Aktivierung einer zusammengesetzten Klasse Q_l sich aus der Summe der Softmax-Aktivierungen der zugeordneten Basisklassen y_q ergibt (siehe Abbildung 4, rechts). Folglich lautet die erweiterte hierarchische Softmax Kreuzentropie

$$E(w) = -\sum_{n=1}^{N}\left(\sum_{k=1}^{K} t_{nk}\ln\left(y_k(\mathbf{x_n}, \mathbf{w})\right) + \sum_{l=1}^{L} t_{nQl}\ln\left(\sum_{q\in Ql} y_q(\mathbf{x_n}, \mathbf{w})\right)\right), \quad \text{Gl. 3}$$

wobei die Indikatorvariablen t_{nk} und t_{nQl} eins sind, falls das n-te Sample zur k-ten Basisklasse beziehungsweise zur Q_l-ten zusammengesetzten Klasse gehört, und null anderenfalls. Auf diese Weise können die zusammengesetzten Klassen nicht vom gelernten Klassifikator prädiziert werden, jedoch während des Trainings als Positivbeispiele für die jeweils zugeordneten Basisklassen und damit implizit als Negativbeispiele für die nicht zugeordneten Basisklassen fungieren.

Für den in Abschnitt 3 vorgestellten Datensatz heißt dies, dass es drei Basisklassen und zwei zusammengesetzte Klassen gibt. Die Basisklassen umfassen jeweils Zeiträume von 50 Jahren und die zusammengesetzten Klassen entsprechend 100 Jahre. Samples, die aufgrund ihrer Epochenbeschreibung einer der zusammengesetzten Klassen zugeordnet werden, können auf diese Weise ins Training einbezogen werden. Ohne die Erweiterung wäre dies nicht möglich gewesen. Ein weiteres Problem beim Trainieren des Netzes stellt neben den uneindeutigen Klassenlabels zum einen die ungleiche Verteilung der Bilder auf die einzelnen Klassen und zum anderen die verhältnismäßig geringe Menge an Trainingsdaten dar. Daher wird als Vorprozessierung eine Augmentierung des hier verwendeten Datensatzes realisiert, um eine repräsentative Grundlage für das Training zu schaffen und eine Überanpassung an die Trainingsdaten zu vermeiden.
Um eine Invarianz gegenüber der Ausrichtung des Objektes gegenüber der Kamera zu schaffen werden die Bilder zunächst um einen zufälligen kleinen Winkel rotiert und anschließend erneut mit einer 50 %-igen Wahrscheinlichkeit um 90°. Eine zusätzliche Variation erfolgt durch zufälliges horizontales und vertikales Spiegeln der Bilder. Anschließend wird ein zufälliger Bereich jedes Bildes ausgeschnitten, der 30 % - 90 % des ursprünglichen Bildes einnimmt. Dieser Ausschnitt wird dann auf die Eingangsgröße des Netzwerkes skaliert. Zusätzlich werden zufällige kleine horizontale und vertikale Scherungen durchgeführt, dies ist für den verwendeten Datensatz sinnvoll, um Deformationen des Musters durch Falten in einigen Stoffen nachzuahmen. Zusätzlich werden Kontrast, Helligkeit, Farbton und Sättigung der Bilder zufällig verändert, da angenommen wird, dass diese durch das Fotografieren der Werke mit unkontrollierten Lichtverhältnissen sowie unterschiedlichem Hintergrund variieren.

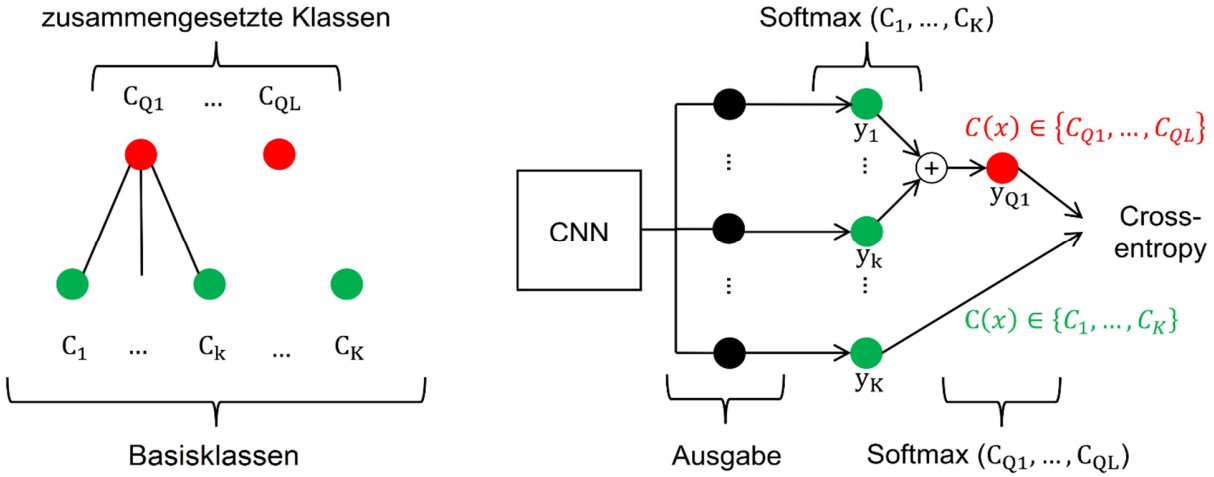

Abb. 4: Schematische Darstellung der Klassenstruktur bestehend aus Basisklassen C_1, …, C_K und zusammengesetzten Klassen (links) sowie dem zugehörigen Ablaufschema der erweiterten Softmax Kreuzentropie (rechts)

6 Experimente

Für das Training der Klassifikationsschicht wird der Datensatz in einen 80 % der Daten umfassenden Trainingsdatensatz sowie einen Validierungs- und einen Testdatensatz, die jeweils 10 % der Daten umfassen, aufgeteilt. Der Validierungsdatensatz, mit dem die optimale Anzahl der Ite-

rationen sowie die Größe der Mini-Batches bestimmt werden variiert im Rahmen der durchgeführten 10-fachen Kreuzvalidierung. Die Parameter der Klassifikationslayer werden mittels dem Adaptive-Momentum (KINGMA & BA 2014) unter Verwendung der Standardparameter optimiert. Dabei wird eine ausgeglichene Klassenverteilung in den Trainingsbatches erzielt, indem für jede Klasse etwa gleich viele Bilder zufällig aus dem Trainingsdatensatz gezogen werden. Hierfür werden für die unterrepräsentierten Klassen mehr synthetische Bilder generiert als für die dominante Klasse "1. Hälfte 20. Jh.", sodass anschließend eine ausgeglichene Verteilung über alle Klassen vorliegt (siehe Tabelle 2).

Tab. 2: Übersicht über die Anzahl der synthetischen Samples pro Klasse.

Klassenbezeichnung	2. Hälfte 19. Jh.	1. Hälfte 20. Jh.	2. Hälfte 20. Jh.
Anzahl originaler Bilder	213	2035	628
Anzahl synthetischer Samples je Bild	964	100	326
Resultierende Anzahl an Samples	205545	205535	205356

Abbildung 5 zeigt exemplarisch ein Originalbild und zugehörige synthetische Variationen.

Abb. 5: Beispiel für ein Bild aus dem (GARÍN) Datensatz (links) und zugehörige synthetische Daten

Die Anzahl der notwendigen Trainingsiterationen wird in jedem Experiment auf Basis des Validierungsdatensatzes ermittelt. Hierfür werden die Gesamtgenauigkeit (Anteil der korrekten Prädiktionen) und die über alle Klassen gemittelten F1-Scores (harmonisches Mittel aus Vollständigkeit und Korrektheit) betrachtet. Ist in dem Verlauf der auf dem Validierungsdatensatz basierenden Qualitätsmaße eine Sättigung zu beobachten, so wird der Testdatensatz evaluiert.
Zunächst wird das Training des Netzes ausschließlich auf Basis der Basisklassen und ohne Datenaugmentierung durchgeführt (BASIS). Somit kann die Prädiktion der Epoche durch den CNN-basierten Klassifikator im Allgemeinen bewertet werden. In den weiterführenden Experimenten wird dann der Einfluss der vorgenommenen Modifikationen evaluiert. Die erste Modifikation ist die Hinzunahme von synthetischen Daten im Training (AUG). Für den Validierungs- und den Testdatensatz werden ausschließlich die originalen Bilder verwendet. Als zweite Modifikation werden statt der synthetischen Samples die Samples der zusammengesetzten Klassen (siehe Tabelle 1) in Kombination mit der erweiterten Softmax Kreuzentropie in (siehe Gleichung 3) hinzugenommen (ERW). Das abschließende Experiment nutzt im Training sowohl synthetische Bilder als auch (ausschließlich originale) Bilder der übergreifenden Klassen (AUG+ERW). Die resultierenden Ergebnisse sind in den Tabelle 3 aufgeführt.

Tab. 3: Zusammenstellung der erzielten Ergebnisse in den einzelnen Experimenten.

Experiment	F1-Score 2. Hälfte 19. Jh. [%]	F1-Score 1. Hälfte 20. Jh. [%]	F1-Score 2. Hälfte 20. Jh. [%]	Gesamt-genauigkeit [%]	F1-Score (Mittel) [%]
BASIS	29,9	79,3	60,9	70,6	56,7
AUG	19,7	83,2	62,3	74,5	55,1
ERW	**32,1**	82,2	**65,6**	74,0	**60,0**
AUG+ERW	19,0	**85,2**	64,0	**77,1**	56,1

Basierend auf einer 10-fachen Kreuzvalidierung wurde hinsichtlich der Gesamtgenauigkeit das beste Ergebnis durch die Verwendung zusätzlicher Daten sowohl durch die Datenaugmentierung als auch durch die erweiterte Verlustfunktion erzielt (AUG+ERW) (siehe Tabelle 3). Die Erweiterung des Trainingsdatensatzes mittels nur einer der beiden Methoden führt zu etwa identischen Ergebnissen auf dem Testdatensatz in Bezug auf die Gesamtgenauigkeit.

Da im verwendeten Datensatz jedoch eine unausgeglichene Verteilung der Bilder auf die drei Klassen vorliegt, ist eine Bewertung der Experimente auf Basis klassenspezifischer Indizes sinnvoll. In Tabelle 4 sind die F1-Scores der einzelnen Klassen aufgeführt. Allgemein fällt auf, dass die Klasse „2. Hälfte 19. Jh." die schlechtesten Genauigkeiten erzielt, was sich auf die geringe Menge an Trainingssamples in dieser Klasse (siehe Tabelle 1) zurückführen lässt. Entsprechend sind für die Klasse „1. Hälfte 20. Jh." mit den meisten Trainingsdaten die höchsten F1-Scores zu beobachten.

Bezüglich der Bewertung der einzelnen Experimente kann festgestellt werden, dass durch die Hinzunahme von synthetischen Trainingssamples die Genauigkeit der am stärksten unterrepräsentierten Klasse deutlich geringer wird; die Genauigkeiten der anderen Klassen können jedoch leicht verbessert werden. Als Ursache wird vermutet, dass im verwendeten Datensatz nicht ausreichend Bilder vorhanden sind, um die Charakteristika der Stoffe aus der Zeit von 1850 bis 1900 zu beschreiben.

Dies wird durch die Verbesserung des entsprechenden F1-Scores bei der Einführung von Zusatzinformationen im Training durch die Hierarchische Verlustfunktion bestätigt. Durch die Hinzunahme von 302 Bilder von Seidenstoffen, deren Herstellungsepoche teilweise in die zweite Hälfte des 19. Jahrhunderts fällt, kann der F1-Score um etwa 2 % verbessert werden. Ebenso kann die Genauigkeit der Klasse „2. Hälfte 20. Jh.", für welche ebenfalls verhältnismäßig wenig Trainingsdaten vorliegen, die Berücksichtigung von Samples mit uneindeutigem Klassenlabel im Training gesteigert werden. Somit erreicht das Experiment mit der Hierarchischen Verlustfunktion den besten mittleren F1-Score von 60 %. Werden zusätzlich zu den Bildern mit uneindeutigem Klassenlabel synthetische Samples zum Training der Klassifikationslayer genutzt, wird ein mittlerer F1-Score vergleichbar mit dem im Bezugsexperiment erreicht. Die Klassen mit verhältnismäßig vielen Trainingsbeispielen weisen verbesserte Genauigkeiten auf, wohingegen für die Klasse „2. Hälfte 19. Jh." wie auch bei dem anderen Experiment mit synthetischen Trainingsdaten einen deutlich schlechteren F1-Score erzielt.

Eine klare Aussage darüber, welches Experiment die besten Ergebnisse liefert, kann nicht getroffen werden, da dies von dem gewünschten Ziel abhängt. Um eine höchstmögliche Gesamtgenauigkeit zu erzielen ist die Kombination aus Datenaugmentierung und erweiterter Verlustfunktion die beste Wahl unter den untersuchten Ansätzen. Ist ein hoher F1-Score der stark

unterrepräsentierten Klasse erwünscht, liefert die erweiterte Verlustfunktion ohne Augmentierung die besten Ergebnisse.

7 Zusammenfassung und Ausblick

In diesem Beitrag wurde zum Umgang mit uneindeutigen Klassenlabels eine erweiterte Verlustfunktion eingeführt, welche Samples mit derartigen Labeln im Training berücksichtigen können. Dies führte in den Experimenten zu einer Verbesserung der klassenspezifischen Genauigkeiten, insbesondere der unterrepräsentierten Klasse. Zudem wurden zur Vergrößerung des Datensatzes und zum Umgang mit einer ungleichen Verteilung der Daten auf die Klassen synthetische Trainingsdaten generiert, wobei die Anzahl der Datenaugmentierungen pro Sample derart gewählt wurde, dass im Trainingsdatensatz eine ausgeglichene Klassenverteilung vorliegt. Die Ergebnisse zeigen, dass die am stärksten unterrepräsentierte Klasse dadurch schlechter prädiziert wird. Da als Ursache die geringe Menge an Daten für die Klasse „2. Hälfte 19. Jh." vermutet wird, wäre in künftigen Untersuchungen eine Erweiterung des Datensatzes durch existierende Sammlungen von Museen von Interesse. Zum einen wären die einzelnen Klassen vermutlich besser durch die Daten repräsentiert und zum anderen könnten durch die größere Menge an Daten einige vortraininerte Layer der Merkmalsextraktion für die neue Aufgabe adaptiert werden. Außerdem könnten im Rahmen von „Multi-Task Learning" Informationen über den Hersteller oder den Herstellungsort für die Prädiktion der Herstellungsepoche einen Mehrwert liefern. Interessant wäre auch die Anwendung der Hierarchischen Verlustfunktion auf andere Datensätze und deren Einfluss auf die Klassifikationsergebnisse. Denkbar wären semantische Hierarchien von Pflanzen- oder Tierbezeichnungen. Zudem wäre in weiterführenden Arbeiten die Entwicklung eines Klassifikators von Interesse, der automatisch eine zusammengesetzte Klasse prädiziert, falls die Prädiktion der Basisklassen aufgrund von fehlenden Informationen nicht möglich ist.

8 Danksagung

The research leading to these results is in the frame of the "SILKNOW. Silk heritage in the Knowledge Society: from punched cards to big data, deep learning and visual/tangible simulations" project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 769504.
All image data as well as the according descriptions were used in this paper's experiments with the friendly permission of Garín 1820 (http://garin1820.com/).

9 Literaturverzeichnis

AYUB, R., ORBAN, C. & MUKUND, V., 2017: Art Appraisal Using Convolutional Neural Networks. http://cs229.stanford.edu/proj2017/final-reports/5229686.pdf

BAR, Y., LEVY, N. & WOLF, L., 2014: Classification of artistic styles using binarized features derived from a Deep Neural Network. Workshop at the European Conference on Computer Vision, Springer, Cham, 71-84.

BLESSING, A. & WEN, K., 2010. Using machine learning for identification of art paintings. Technical Report, Stanford University, 5 pages. https://pdfs.semanticscholar.org/1d73/0a452a5c03cc23f90d4fde71c08864f31c35.pdf

BISHOP, C., 2006: Pattern Recognition and Machine Learning. 1st edition, Springer, New York, 235-240.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. & FEI-FEI, L., 2009: ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 248-255.

GARÍN: Image data and according descriptions with friendly permission of Garín 1820. URL: http://garin1820.com/

HENTSCHEL, C., WIRADARMA, T. & SACK, H., 2016: Fine tuning CNNS with scarce training data - Adapting imagenet to art epoch classification. IEEE International Conference on Image Processing (ICIP), 3693-3697.

HE, K., ZHANG, X., REN, S. & SUN, J., 2016a: Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition, 770-778.

HE, K., ZHANG, X., REN, S. & SUN, J., 2016b: Identity mappings in deep residual networks. European conference on computer vision, Springer, Cham, 630-645.

HICSONMEZ, S., SAMET, N., SENER, F. & DUYGULU, P., 2017: DRAW: Deep networks for Recognizing styles of Artists Who illustrate children's books. ACM on International Conference on Multimedia Retrieval, 338-346.

KINGMA, D. P. & BA, J. L. 2014: Adam: A method for stochastic optimization. arXiv preprint, arXiv:1412.6980.

KRIZHEVSKY, A., SUTSKEVER, I. & HINTON, G. E., 2012: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, **25**(1), 1097-1105.

RAZAVIAN, A., AZIZPOUR, H., SULLIVAN, J. & CARLSSON, S., 2014: CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. IEEE conference on computer vision and pattern recognition workshops, 806-813.

RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C. & FEI-FEI, L., 2015: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, **115**(3), 211-252.

SALEH, B. & ELGAMMAL, A., 2016: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. International Journal for Digital Art History, 70-93.

SIMONYAN, K. & ZISSERMAN, A., 2014: Very deep convolutional networks for large-scale image recognition. arXiv preprint, arXiv:1409.1556.

SUR, D. & BLAINE, E., 2017: Cross-Depiction Transfer Learning for Art Classification. Technical report for CS 231A and CS 231N at Stanford.

TAN, W. R., CHAN, C. S., AGUIRRE, H. E. & TANAKA, K., 2016: Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. IEEE International Conference on Image Processing (ICIP), 3703-3707.

VURAL, V. & DY J. G., 2016: A hierarchical method for multi-class support vector machines. In: Proceedings of the 21st International Conference on Machine Learning, ACM, 105.

WU, Z. & SAITO, S., 2017: HiNet: Hierarchical Classification with Neural Network. 5th International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1705.11105.

XIAO, Z., LIU, X., WU, J., GENG, L., SUN, Y., ZHANG, F. & TONG, J., 2018: Knitted fabric structure recognition based on deep learning. The Journal of the Textile Institute, 1-7.

YOSINSKI, J., CLUNE, J., BENGIO, Y. & LIPSON, H., 2014: How transferable are features in deep neural networks? Advances in Neural Information Processing Systems 27, 3320-3328.