

Masterarbeit

Layout-Erkennung in digitalisierten Dokumenten mittels Neuronaler Netzwerke

von Jakob Schmolling

Matrikelnummer:

Datum: 26. April 2018

Fachbereich 4 Wirtschaftswissenschaften II
Internationale Medieninformatik (M.A.)

Erstgutachter: Prof. Dr. Klaus Jung
Zweitgutachter: Prof. Dr. Kai-Uwe Barthel

Inhaltsverzeichnis

1 Digitalisierte Dokumente	7
1.1 Dokumente	7
1.2 Digitale Bibliotheken	8
1.3 Datenrepresentation	8
1.4 Schritte in der Verarbeitung von Dokumentenbildern	9
2 Theoretische Grundlagen	11
2.1 Maschinelles Lernen	11
2.2 Künstliche Neuronale Netzwerke	12
2.3 Aktivierungsfunktionen	13
2.4 SGD	14
2.5 Xavier initialization	14
2.6 Regularisierung	14
2.7 CNN	14
2.8 Vorverarbeitung	15
2.9 SLIC Superpixel	15
3 Reproduktion bisheriger Ergebnisse	17
3.1 Andere Ansätze	17
3.2 Datensatz	17
3.3 Auswahl und Beschreibung des Datensatzes	17
3.4 DIVA-HisDB	18
3.5 Kai Chen, Seuret u. a. (2017)	20
3.6 Netzwerk-Architektur	20
3.7 Training	21
3.8 Nachverarbeitung	21
3.9 Xu u. a. (2017)	21
3.10 VGG	21
3.11 Deconvolution	21
3.12 Ergebnisse	21
4 Selbstüberwachtes Lernen	23
4.1 Unüberwachtes Lernen von unterscheidbaren Features	24
5 Umsetzung	25
5.1 Chen	25
5.2 Xu	25
5.3 Discrimitve	27
5.4 Evaluierung	27

5.5 Fazit	28
Literaturverzeichnis	29

Einleitung

Diese Arbeit beschäftigt sich mit der Layout-Segmentierung von Digitalisierten Dokumenten mittels Neuronaler Netzwerke.

Kapitel 1 beschreibt die Motivation für die Dokumentensegmentierung und aktuelle Entwicklungen im Bereich der Dokumentendigitalisierung.

Kapitel 2 beschreibt die Theorie der Künstlichen Neuronalen Netzwerke.

Kapitel 3 setzt sich mit den mit zwei Forschungsergebnissen der Dokumentensegmentierung auseinander indem versucht werden zusle

Kapitel 4 erläutert die Methode des selbstüberwachten Lernen. Die Methode wird auf die reproduzierten Forschungen angewendet.

Kapitel 5 erläutert Details, Probleme und Ergebnisse der umgesetzten Reproduktionen und dem Einsatz von selbstüberwachten Lernen.

1 Digitalisierte Dokumente

Menschen erstellen Dokumente schon seit mehr als 4000 Jahren (Smith 2014, S. 13). Dokumente, von Tontafeln (siehe Abb. 1.1) bis hin zu rein digitalen Dokumenten, ermöglichen Kommunikation über zeitliche und örtliche Grenzen hinweg. Diese Technologie ist aus unserer modernen Kultur nicht mehr wegzudenken. Seit den 50er Jahren begann die Forschung im Bereich der optischen Zeichenerkennung (engl. OCR)(David S. Doermann 2014). OCR fand zuerst Einsatz in genau spezifizierten Problembereichen zum Beispiel die Erkennung von Druckbuchstaben einer Schreibmaschine. Je mehr Dokumente digitalisiert wurden, desto klarer wurde es das Dokumente mehr als eine Kette von Zeichen sind. Information können in Dokumenten über die Position der Zeichen und Skalierung von Zeichen vermittelt werden. Zum anderen bestehen Dokumente aus Inhalten die semantische Bedeutung haben, aber nicht als Zeichenkette codiert werden können.

Wirtschaftsinteressen trieben die Entwicklungen von Dokumentenverarbeitungssystemen in einigen Bereichen sehr weit voran, wie zum Beispiel bei der Verarbeitung von Geschäftsbriefen und Formularauswertung. Eine Spezialisierung auf bestimmte Dokumentenklassen ist immer noch eine Notwendigkeit angesichts der unzähligen, veränderbaren und nicht fest gelegten Gestaltungsmöglichkeiten für Dokumente (H. S. Baird und Tombre 2014, S. 69).

1.1 Dokumente

Typischerweise beschreiben wir ein Dokument als ein Papier mit einer Nachricht darauf. Diese Nachricht kann textueller Art sein, dass bedeutet Glyphen die in horizontalen oder vertikalen Linien angeordnet sind (je nach Sprache)(P. H. S. Baird 2014). Diese Textlinien sind dann meist in Textblöcken organisiert. Dokumente können auch grafische Inhalte haben. Dokumente die mit modernen maschinellen Verfahren hergestellt sind haben eine einheitlichere Form in der Hinsicht, dass sie nicht nur einen einheitlichen Schriftsatz besitzen sondern auch alle anderen typografischen Parameter sind einheitlicher, auch Textlinien sind gerade und parallel.



Abbildung 1.1: Königsliste (2047 Jahre v.Chr. ebd.)

Warum sind alte Dokumente noch schwerer?

1.2 Digitale Bibliotheken

Immer mehr Bibliotheken und Archive arbeiten daran Sammlungen von Dokumenten zu digitalisieren und im Internet zur Verfügung zu stellen. Digitale Sammlungen sind dann nicht mehr an die Restriktionen von analogen Dokumenten gebunden und können deshalb auch seltene Dokumente oder Einzelstücke einer großen Menge von Nutzern zugänglich machen. Dies erleichtert besonders die Forschung an historischen Dokumenten, da diese oft so wertvoll und fragil sind, dass die Lagerung und Nutzung der Originale strengen Auflagen unterliegen.

Was ist eine digitale Bibliothek?

Die Digitalisierung ist aber nicht kostenfrei. Die Kosten belaufen sich bei Digitalisierungsprojekten in Deutschland auf etwa 10 bis 50 Cent pro Seite (Opitz und Stäcker 2009). Die komplette Erfassung kann nochmals teurer sein, da die Erfassung von Strukturdaten zusätzliche Kosten verursacht. Im Digitalisierungsprojekt *dünnhaupt digital* ist die Strukturdatenerfassung mit 40 Cent pro Seite der größte Ausgabenpunkt. *ebd.* bemerkt auch, dass die Volltextgenerierung mittels OCR nicht für eine sinnvolle Onlinenutzung nicht ausreichend ist. Erst die Erfassung von Seitenstrukturen erlaubt “einen gezielten Zugriff auf logische Texteinheiten” (*ebd.*, S. 372).

Anforderungen an digitale Dokumente?

Gescannnte Dokumente ohne eine symbolische Kodierung laufen Gefahr irrelevant zu werden, da ihre Handhabung viel schlechter ist als die von “purely digital information” (H. S. Baird 2003, S. 10).

1.3 Datenrepresentation

Nichtdestotrotz sind enorm viele digitalisierte Dokumente im Internet verfügbar. Die größten Sammlungen sind unter anderem Google Books und das Internet Archive. Diese Sammlungen haben das Ziel möglichst viele Dokumente bereitzustellen. Die Qualität kann dabei stark variieren.

Andere Sammlungen beschränken sich auf einen klarer definierten Korpus und Digitalisierungsmethoden.

Die digitalisierte Sammlungen der Staatsbibliothek zu Berlin (Staatsbibliothek zu Berlin 2016) umfasst 134645 Werke aus dem Bestand der Bibliothek.

research: e-codices Das e-codices Projekt beschäftigt sich seit 2005 mit der Digitalisierung von mittelalterlichen und neuzeitlichen Handschriften. Aktuell befinden sich 1947 Handschriften in der Sammlung (e-codices 2018). Die Dokumente haben eine Auflösung von mindestens 300ppi und eine Farbtiefe von 16bit.

Vereinheitlichung

1.4 Schritte in der Verarbeitung von Dokumentenbildern

Die Dokumentensegmentierung ist ein Vorverarbeitungsschritt für weitere Schritte der Dokumentenverarbeitung.

2 Theoretische Grundlagen

Künstliche Neuronale Netzwerke (kurz NN) dominieren den Forschungsbereich der Bilderkennung. Besonders die Klasse der faltenenden Neuronalen Netzwerks (engl.: *convolutional neural network*), kurz CNN, konnte viele Erfolge für sich beanspruchen. Rekordbrechende Ergebnisse bei der Klassifizierung von handgeschriebenen Ziffern (LeCun, Boser u. a. 1989) und des ImageNet-Wettbewerbs (Krizhevsky, Sutskever und G. E. Hinton 2012) motivieren dazu CNN-Methoden auch in anderen Bereichen anzuwenden.

Neuronale Netzwerke benötigen weniger Entwicklungsaufwand und können zudem einfacher von wachsenden Rechenkapazitäten und Datenmengen gebrauch machen (LeCun, Bengio und G. Hinton 2015, S. 436). Aber die Wissenschaft der NN ist ein Feld, dass durch die aktuelle Praxis mehr als durch Theorie geprägt ist. Das bedeutet zum einen das theoretische Grundlagen noch nicht gefestigt sind. Zum anderen ist die Theorie nur eine Faktor für den erfolgreichen Einsatz von NNs. Je mehr Parameter unklar sind desto schwieriger wird eine Reproduktion. Die Reproduktion von Ergebnissen ist aber ein wichtiger Bestandteil in jedem Forschungsgebiet.

2.1 Maschinelles Lernen

Neuronale Netzwerke und CNN sind Beispiele für Maschinenlernalgorithmen (engl.: *maschine learning algorithms*). Mitchell 1997 definiert einen Maschinenlernalgorithmus wie folgt: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P improves with experience E”.

Die Erfahrung E kann zum Beispiel ein Datensatz X mit zugehörigen Klassen Y sein. Im Fall der Dokumentensegmentierung kann ein Element x_i ein Bildausschnitt sein, dessen Kategorisierung ist dann das Label y_i (siehe Abb. 2.1).

Die Aufgabe T ist in den meisten Fällen eine statistische Modellierung. Es wird angenommen, dass X und gegebenfalls Y mithilfe einer Wahrscheinlichkeitsfunktion p modelliert werden können. Ein unüberwachter Lernalgorithmus versucht die Verteilung $p(x)$ direkt

Warum NN?

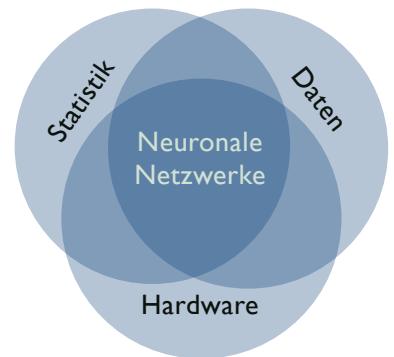


Abbildung 2.1: Diziplinen im Bereich Neuronale Netzwerke

Unterschied replikation, reproduktion?

zu lernen. Beispiele für unüberwachte Algorithmen sind der k-Means-Algorithmus oder der PCA-Algorithmus.

Bei überwachten Lernalgorithmen wird versucht die Wahrscheinlichkeitsverteilung $p(y|x)$ implizit zu “lernen”. Der Lernalgorithmus soll dabei eine Funktion finden $f : \mathbb{R}^n \rightarrow \{1, \dots, k\}$ so das $y = f(x)$ (Goodfellow, Bengio und Courville 2016, 97 ff). Ein Lernalgorithmus der diese Funktion findet ist ein überwachter Lernalgorithmus (engl.: *supervised learning algorithm*) Die Trennung zwischen überwacht und unüberwacht ist nur eine grobe Einteilung. Tatsächlich sind die Übergänge fließend. Manche Maschinenlernverfahren benutzen beide Methoden.

Die Performanz P der Vorhersagefunktion f wird im einfachsten Fall mit der Genauigkeit, dem Verhältniss von richtigen Vorhersage zu Falschen, gemessen. Tabelle 2.1

		y	
		Positiv	Negativ
\hat{y}	Positiv	true positive	false positive
	Negativ	false negative	true negative

Tabelle 2.1: Wahrheitstabelle

2.2 Künstliche Neuronale Netzwerke

Künstliche Neuronale Netzwerke (kurz.: KNN) sind eine Modellierungsmethode des statistischen Lernens, die von der Natur inspiriert wurde. Das Neuron ist der Grundbaustein eines Netzwerkes und verarbeitet eingehende Signale. Ein KNN besteht meistens aus zwei Teilen: einer linearen Funktion über die Eingabesignale und einer nicht-linearen Aktivierungsfunktion.

Zur besseren Erläuterung der Theorie soll hier ein Teilproblem der Dokumentensegmentierung dienen. In diesem Fall besteht der Datensatz X aus Bildausschnitten die im Bildzentrum Text enthalten. Wenn ein Element x_i Text enthält, dann ist das $y_i = 1$ andernfalls 0.

X	Y
0	0
0	0
foo	1
bar	1

Abbildung 2.2: Bildausschnitte mit und ohne Textinhalt

$$f \left(\begin{matrix} \text{bar} \end{matrix} \right) = 1 \quad (2.1)$$

Wir wollen eine Funktion finden, die mit hoher Wahrscheinlichkeit eine Vorhersage

$$\theta_{ML} = \arg \max_{\theta} P(Y|X, \theta) \quad (2.2)$$

Das einfachste Beispiel ist die Klasse der vollvernetzen KNN. Jedes Element in einem Eingangssignalvektor x wird mit einem Parameter aus dem Gewichtsvektor w berechnet. Dazu kommt der Biaswert b .

$$\hat{y} = w^\top x + b \quad (2.3)$$

$$f(X) = \sum_{m=1}^M g_m (\omega_m^\top X) \quad (2.4)$$

Target als nonlineare Funktion dieser Features modellieren.

2.3 Aktivierungsfunktionen

$$\text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad (2.5)$$

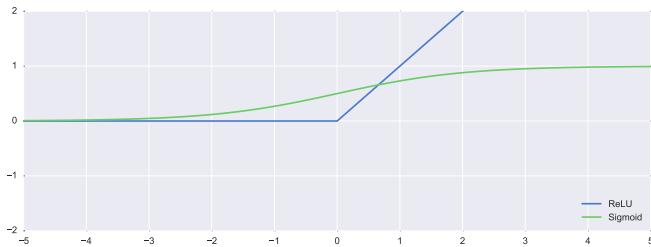


Abbildung 2.3: ReLU- und Sigmoid-Funktionsplot

ReLU

$$f(x) = \max(0, x) \quad (2.6)$$

Softmax

2.4 SGD

2.5 Xavier initialization

2.6 Regularisierung

Durch SGD und in kleinen Datensets entsteht Samplingrauschen (engl.: *sampling noise*). Eine kleine Auswahl von Beispielen kann Zusammenhänge enthalten, die nicht in der tatsächlichen Verteilung der Daten vorhanden ist. Regularisierung können helfen eine Überanpassung des Models zu verhindern.

Dropout

Dropout ist eine Regularisierungsmethode für Neuronale Netzwerke, die versucht durch zufälliges Ausschalten einzelner Aktivierungen Koadaptionen zu verhindern.

Während einer Trainingsiteration wird ein Element mit einer Wahrscheinlichkeit p ausgeschaltet (Aktivierung gleich 0). Dadurch wird bei jedem Trainingsschritt ein anderes Modell trainiert. In der Testphase werden die Gewichte mit p skaliert.

Netzwerke mit Dropout generalisieren besser auf noch nicht gesehene Daten.

Weight decay

2.7 CNN

Die Anzahl der Parameter in der Gewichtsmatrix W wird von der Größe des Inputs bestimmt. Möchte man Bilder ein Bild mit der der CNN kombinieren zwei Konzepte der Bilderverarbeitung: Neuronale Netzwerke und Filter. Klassifikationsprobleme wurde traditionell in zwei Schritten gelöst. Zuerst wurden Featuredeskriptoren entwickelt welche dann als Input für trainierbare Klassifizierer verwendet wurden (Rawat und Z. Wang 2017, S. 2353).

So können zum Beispiel Kanten in einem Bild ein Klassifizierungsgrundbaustein sein. Um die Kanten in einem Bild zu finden wird das Bild mit einem Kernel gefaltet. Der Begriff Faltung wird üblicherweise für die Verknüpfung von zwei realwertigen Funktionen verwendet. Die analoge Form der Faltung ist eine lineare, translationsinvariante Funktion, siehe Süße und Rodner 2014, S. 28.

Nachteile fully connected NN für Bilderverarbeitung

Sind Filter vorläufer von CNN?

Was ist convolutional?

$$\text{convolutional} \quad (2.7)$$

Im Bereich der Bilderverarbeitung wird schon länger eine abgeänderte Faltungsfunktion genutzt: die Autokorrelation (engl.: *cross-correlation*). Bild und Filter sind 2D-Funktionen die zum einen nur für diskrete Intervale definiert sind zum anderen für Bereiche aushalb des ihres Definitionsbereichs gleich 0 sind. Zudem wird überlicherweise der Index für den Kernel addiert (siehe Gleichung (2.7)).

$$I * K(i, j) = \quad (2.8)$$

Ein CNN ist eine Neuronales Netzwerk, dass in mindesten einen Layer die Faltung anstatt der normalen Matrixmultiplikation verwendet (Goodfellow, Bengio und Courville 2016, S. 321). ebd. sieht drei Vorteile von CNN gegenüber voll-vernetzten NN: verringerte Konnektivität, gemeinsame Parameternutzung, eqvariante Darstellung. Wie schon erwähnt steigt die Anzahl der Parameter in einem voll-vernetzten NN je größer der Input wird, weil jedes Neuron mit jeder Aktivierung der vorhergehenden Schicht verbunden ist. Durch die Faltungsoperation erhält jedes Neuron nur die Signale die im Bereich des Filterkernels liegen. Dieser Bereich wird auch rezeptives Feld genannt. Durch die Schichtung von mehreren Faltungen vergrößert sich das rezeptive Feld der Neuronen in tieferen Schichten (siehe Abschnitt 2.7).

2.8 Vorverarbeitung

Für die Verarbeitung von Dokumentenseiten werden meistens noch weitere Algorithmen eingesetzt um Klassierung und Segementierung zu erleichtern. Da wir nur an Inhalten interessiert sind die auf ein Medium aufgetragen wurden (z.B. Tinte auf Papier) ist eine Binarisierung des Bildes ein hilfreicher Schritt. Eine Binarisierungsfunktion $bin(I) : f : i \in \{0 \dots 255\} \rightarrow b \in 0, 1$

2.9 SLIC Superpixel

Der SLIC-Algorithmus basiert auf dem k-Means-Algorithmus und teilt Pixel innerhalb eines 5D-Raums in Cluster ein. In jedem Arbeitsschritt werden Pixel dem Clusterzentrum mit der geringsten Distanz zugeordnet und danach werden die Clusterzentren neu berechnet. Das Distanzmaß D_s zu den Clusterzentren $k = [1, K]$ basiert auf den Farbabstand im Lab-Farbraum d_{lab} und den räumlichen Abstand d_{xy} :

$$d_{lab} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \quad (2.9)$$

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \quad (2.10)$$

$$D_s = d_{lab} + \frac{m}{S} d_{xy} \quad (2.11)$$

Drei Hauptvorteile?



Abbildung 2.4: Rezeptives Feld in einem mehrschichtigen Netz

Nachteile CNN?

Filterbeispiel?

Der Faktor m ermöglicht eine Gewichtung der zwei Distanzmaße. Je höher der Faktor desto kompakter werden die Superpixel. Abb. 3.2 zeigt das Ergebniss des Algorithmus mit unterschiedlichen Parameter m angewendet auf eine Dokumentenseite.

3 Reproduktion bisheriger Ergebnisse

What I cannot create, I do
not understand
— Richard Feynman

Ziel Im folgenden werden Paper genauer untersucht, die sich mit dem Problem der Dokumentensegmentierung mit Hilfe von CNNs nähern. Das erste Experiment basiert auf dem neusten Paper von Mitgliedern der DIVA-Gruppe: Kai Chen u. a.

welche Experimente?

Das zweite Experiment basiert auf der Forschung des Gewinners des IDCAR2017-Wettbewerbs: Xu u. a.

Beide Untersuchungen benutzen den gleichen Datensatz als Grundlage für ihre Experimente.

3.1 Andere Ansätze

Wick und Puppe 2017 Im Bereich der Bibliotheswissenschaften besteht ein großes Interesse an Klassifizierung von Buchseiten zur besseren Erschließung. McConaughey, Dai und Bamman 2017 klassifizieren Buchseiten anhand von textbasierten Features in 4 Kategorien.

3.2 Datensatz

3.3 Auswahl und Beschreibung des Datensatzes

DoermannHandbookdocumentimage2014 listen 5 Aspekte die bei der Erstellung von Datensätzen zu beachten sind:

- Auswahl der Daten
- Datenbeschaffung
- Ground Truth Definition
- Ground Truth Annotation

- Speicherformat
- Struktur und Organisation

3.4 DIVA-HisDB

Die DIVA (Document, Image and Voice Analysis) Gruppe der Universität Fribourg hat im Kontext der Forschungsprojekte HisDoc und HisDoc 2.0 das Datenset DIVA-HisDB erstellt. Die HisDoc-Projekte beschäftigen sich mit der automatischen Analyse von historischen Dokumenten und wie man diese für Historiker nutzbar machen kann.

<http://diuf.unifr.ch/main/diva/>

Foteini Simistira u. a. 2016

Für den Datensatz wurden Dokumente mit komplexen Layout aus der Virtuellen Manuskriptbibliothek der Schweiz (<http://www.e-codices.unifr.ch/en>) ausgesucht. research: Manuskriptbibliothek

Die Manuskripte enthalten neben dem Haupttext auch Randnotizen und Text-Dekorationen. Randnotizen befinden sich auch teilweise zwischen den Zeilen des Haupttexts. DIVA-HisDB besteht aus 150 Dokumenten, aufgeteilt in Trainings-, Validierungs- und Testset (siehe Tabelle 3.1). Hinzu kommen 30 Seiten die für die finale Wertung des Wettbewerbs “ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts” verwendet wurden.

Name	Auflösung	Training	Validierung	Test	Test ICDAR 2017)
CB55	4872×6496	20	10	10	10 Tabelle 3.1: Aufteilung der
CSG18	3328×4992	20	10	10	10 Seiten des DIVA-HisDB-
CSG863	3328×4992	20	10	10	10 Datenssets

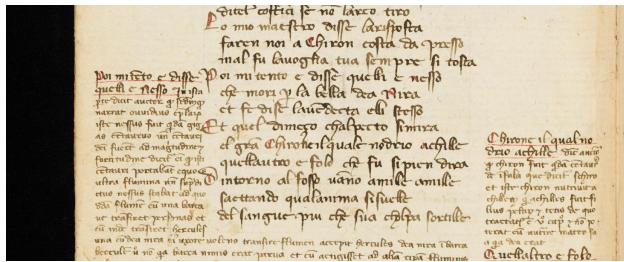
Alle Daten können direkt von der Webseite der DIVA-Gruppe heruntergeladen werden.

Die Manuskripte wurden mit einer Auflösung von 600 dpi gescannt und sind im JPEG-Format gespeichert. Die Tabelle 3.3 zeigt Beispiele aus den drei Datensätzen.

Der Datensatz wurde semi-automatisch mit 3 Annotation (Haupttext, Kommentare, Dekorationen) versehen. Alle Bereiche die nicht annotiert sind werden als Bildhintergrund betrachtet.

Diese Ground-Truth-Annotationen sind im PAGE-XML-Format und als “pixel-label” PNG-Bilder gespeichert. Die Abb. 3.1 zeigt eine Beispieldseite mit den zugehörigen Labels auf Pixelebene.

Die Menge an Pixeln pro Klasse ist sehr unterschiedlich. Die Tabelle 3.2 zeigt das in jedem Dokumentenset der Hintergrund deutlich überwiegt.



(a) Dokumentenbild

Dicit officia se non laresco tiro !
O mio maestro dico sacrificia
farem noi a Chiron costa da messa
inal fu luogofia tua sempre si tosta
Qui mistero e dico. Vi mi teme e disse quel li e nello
quale m'ha fatto tu mori y la bella dea mera
et fe dico l'autetta ehi trese
Et quel domino sacrificato sombra
dei fiume in eterni qual dico e follo che fin s'puen drea
d'eterno per le fiume quell'autetta e follo che fin s'puen drea
altra fumma in fiume. Intorno al foso uano amile omille
faciendo qualanima fumula
del sangue pur che sua cappa fumula
qui ordine mett in agere uelina transire flumen accepit ferula. Da uera fiuma
ferula. E io se buca nonni este pessimi et cu' amiglier ad alia cosa fumula

(b) "pixel-label"

Abbildung 3.1: Hintergrund: weiss, Haupttext: blau, Kommentare: grün, Dekorationen: rot. Tabelle 3.2: Verteilung der Klassen in Prozent(F. Simistira u. a. 2017, S. 1362)

Set	Hintergrund	Kommentar	Dekoration	Text
CB55	82.41	8.36	0.55	8.68
CSG18	85.16	6.78	1.47	6.59
CSG63	77.82	6.35	1.83	14.00

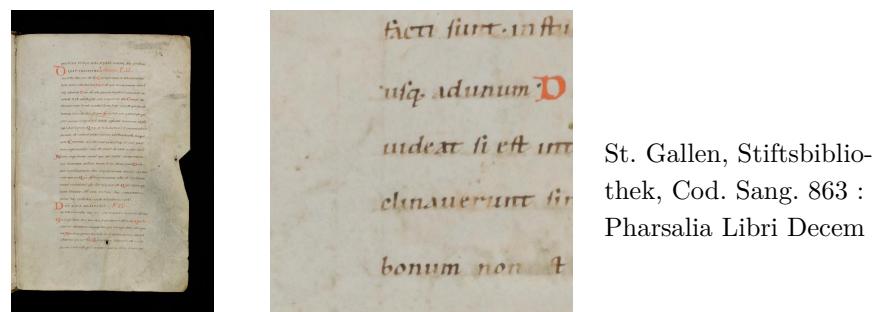
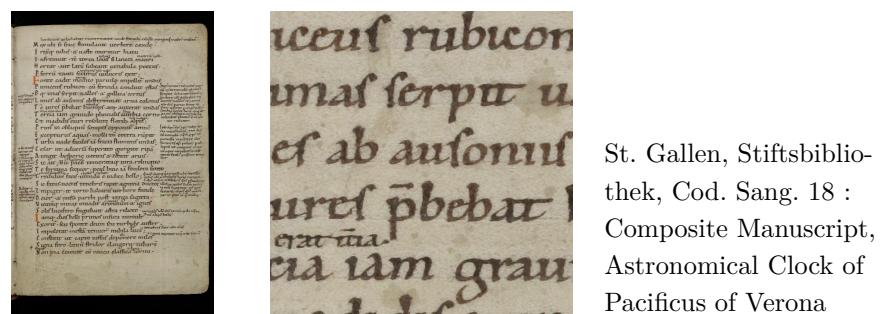


Tabelle 3.3: HisDB Beispiele mit Detailauschnitt

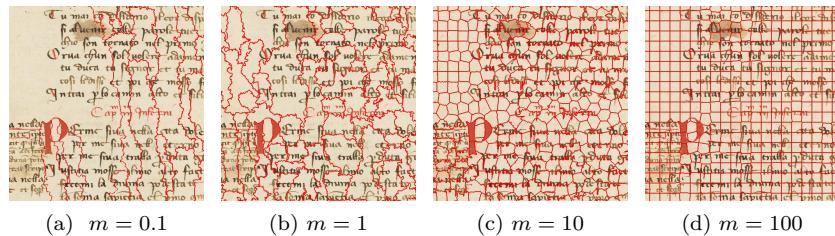
3.5 Kai Chen, Seuret u. a. (2017)

Das Paper „Convolutional Neural Networks for Page Segmentation of Historical Document Images“ von Kai Chen u. a. betrachtet die Dokumentensegmentierung als ein „pixel labeling problem“. Aber eines der größten Probleme bei der Verarbeitung von Dokumentenseiten die Größe der Scanbilder. Genauer gesagt modellieren Kai Chen, Seuret u. a. die Segmentierung als ein Superpixelklassifizierungsproblem. Die Bilder im HisDB-Datenset sind mit einer Auflösung von 4872×6496 wesentlich größer als andere Datensets. Um den Prozess zu beschleunigen werden nicht alle Pixel sondern nur etwa 3000 Pixelcluster klassifiziert.

Welche?

Vorverarbeitung

Kai Chen, Seuret u. a. skalieren alle Bilder mit einem Faktor von 2^{-3} und wenden dann den Superpixelalgorithmus SLIC (simple linear iterative clustering) an (Radhakrishna Achanta u. a. 2010) um die Dokumentenseiten in Superpixel einzuteilen. Ein 28×28 -Bereich um das Zentrum des Superpixel wird dann mithilfe eines CNN klassifiziert. Diese Klassifizierung wird dann allen Pixel innerhalb des Superpixels zugewiesen.



R. Achanta u. a. stellen später zwei wichtige Erweiterungen vor:
Normalisierung des Distanzmaßes und Adaptive-SLIC.

Kai Chen und Seuret nennen keine Details zur Wahl der Superpixel-Parameter, verweisen aber auf eine frühere Studie die sich mit unterschiedlichen Superpixel verfahren beschäftigt (K. Chen u. a. 2016). Die Studie versucht einen Autoencoder (siehe ??) zu trainieren auf Basis von Superpixeln und vergleicht dabei den Einfluss von unterschiedlichen Superpixel-Methoden. Die veröffentlichten Ergebnisse zeigen aber nur die Performanz im Bezug auf die Zahl der Cluster $n \in \{10^3, 50^3, 100^3, 200^3\}$ und dem Skalierungsfaktor $\alpha \in \{2^{-2}, 2^{-3}\}$

Abbildung 3.2: Die SLIC-Pixelgrenzen sind in rot dargestellt

adaptive slic

Problem gt zu treffen

3.6 Netzwerk-Architektur

Kai Chen und Seuret beschreiben die Struktur des CNN als $28 \times 28 \times 1 - 26 \times 26 \times 4 - 100 - M$.

Während des Trainings wird das Ground-truth Label des Zentrumpixels als Label für den Superpixel verwendet.

3.7 Training

3.8 Nachverarbeitung

3.9 Xu u. a. (2017)

Das Paper „Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks“ von Xu u. a. verfolgt einen anderen Ansatz. Xu u. a. verwenden eine Netzwerk

3.10 VGG

3.11 Deconvolution

3.12 Ergebnisse

4 Selbstüberwachtes Lernen

Ein Teil des Erfolges von CNNs ist auf die Verfügbarkeit von großen annotierten Datensätzen wie ImageNet zurückzuführen (Sun u. a. 2017). Jedoch ist die Anzahl der Bilder im ImageNet-Datensatz in den vergangenen Jahren gleich geblieben, während die verfügbare Netzwerkkapazität und GPU-Rechenleistung angestiegen ist. Sun u. a. zeigen mithilfe des JFT-300M-Datensets (300 Millionen annotierte Bilder), dass die Performanz von allen getesteten Aufgaben (Klassifizierung, Segmentierung, etc.) logaritmisch zur Datenmenge ansteigt.

Für einen linearen Performanzanstieg bräuchte man exponentiell mehr Daten. Die Annotierung von Datensets ist immer ein teurer Prozess (Valveny 2014, S. 988). Die manuelle Annotation einer Dokumentenseite kann mehr als eine Stunde in Anspruch nehmen.

Warum besteht Interesse an unsupervised/selfsupervised Lernen?

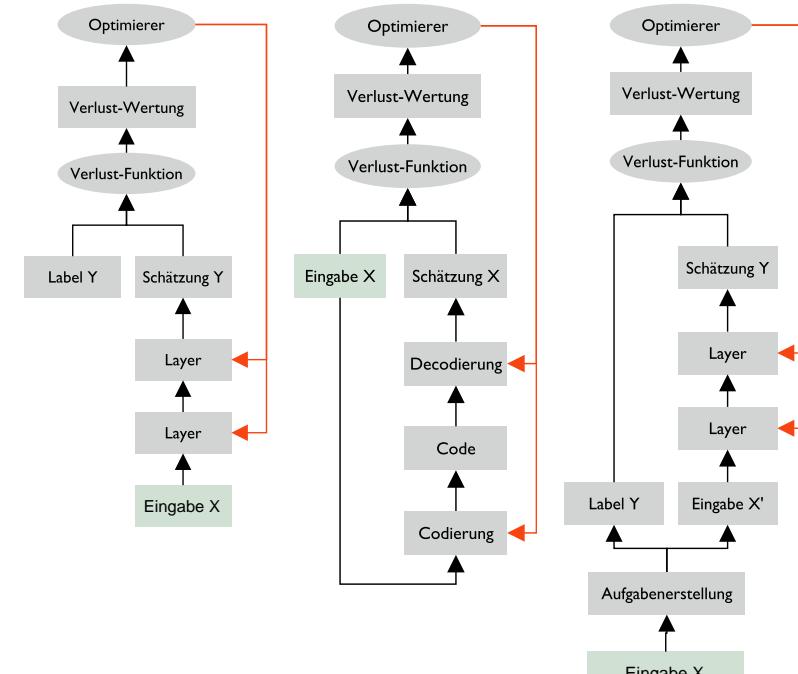


Abbildung 4.1: Schematische Darstellung von ML-Methoden

4.1 Unüberwachtes Lernen von unterscheidbaren Features

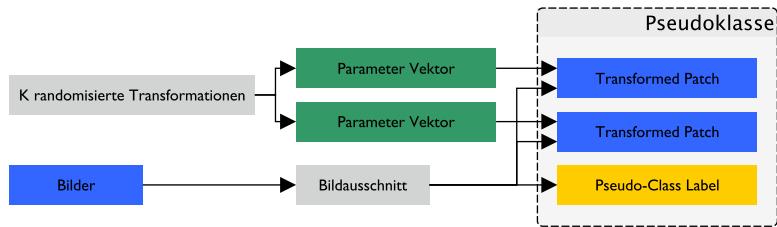


Abbildung 4.2: Workflow

Jigsaw Noroozi und Favaro 2016

Dosovitskiy u. a. 2016 Loss-Funktion skaliert nicht auf viele Klassen.
Letzter Klassifizierungslayer ist Fully-Connected. Je mehr Image-Patches desto mehr Parameter.

Doersch und Zisserman 2017 benutzt Triplet-Loss. Der Verlust wird mit einem Positiv und einem Negativbeispiel berechnet. Die Idee dafür kommt von X. Wang und Gupta. Die Kosinus-Ähnlichkeit misst den Winkel zwischen zwei Vektoren. Vektoren mit einem Winkel von 0 haben eine Kosinus-Maß von 1. Der größtmögliche Winkel hat eine Kosinusmaß von 0.

5 Umsetzung

Nebenziel bei der Umsetzung

Ziel der Arbeit war es nicht nur Ergebnisse zu reproduzieren, sondern auch selbst reproduzierbar zu sein. Dafür sollten alle Prozesse, Datenverarbeitungsschritte und gewählte Parameter in einheitlicher Form definiert sein.

Python und PyTorch

Python ist eine objektorientierte, interpretierte Programmiersprache. Zur Bildverarbeitung wurde die Python-Bibliothek scikitimage verwendet. Die neuronalen Netze wurden mithilfe von PyTorch umgesetzt. PyTorch ist ein Framework zur automatischen Differenzierung von skalarer Funktion (Paszke u. a. 2017).

5.1 Chen

Kai Chen und Seuret nennen nicht die genaue Zahl des Verwendeten Kompaktheitsparameter m (siehe ??). Die Konfiguration der Superpixel setzt aber eine Obergrenze für die maximal erreichbare Genauigkeit. Um den besten Parameter zu finden wurde die Bilder mit der Vorbearbeitungsmethode von Kai Chen und Seuret in Superpixel aufgeteilt. Die Klassifizierung der Pixel wurde direkt aus der ground truth übernommen. Dieser Vorgang wurde für 10 zufällige Bilder des DIVA-Datensatzes wiederholt und ein Durchschnitt gebildet.

aslic? Netz Lernt, aber nur MNIST Boundary pixel werden als GT label verwendet

5.2 Xu

multilabel vorhersage?

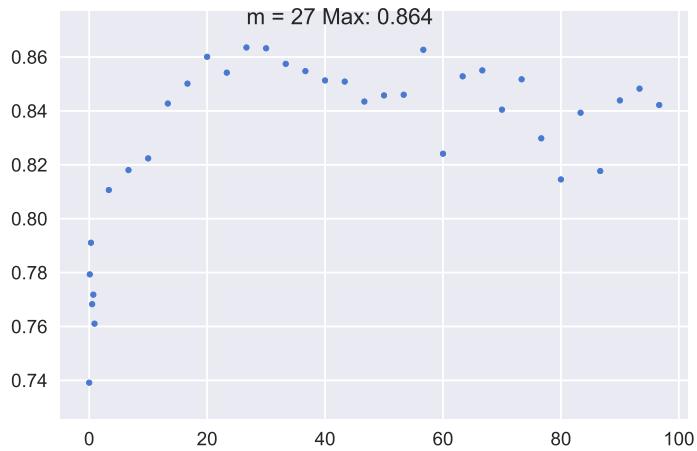


Abbildung 5.1: Maximale Genauigkeit in Abhangigkeit von m

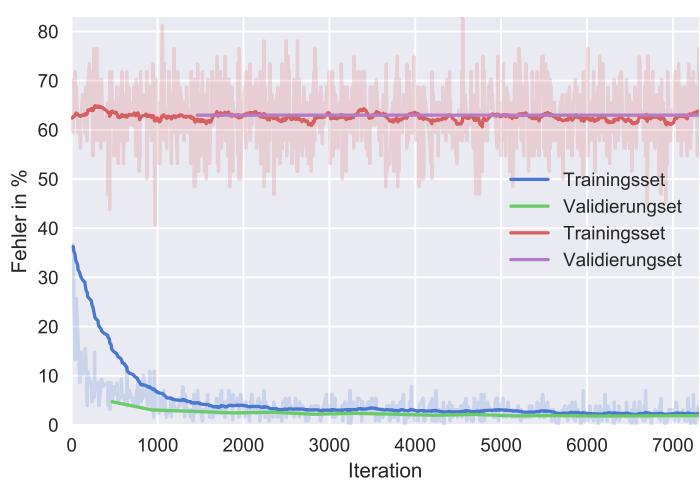


Abbildung 5.2: Fehlerrate CheNet(4,4,0) wahrend der Trainingsiterationen

5.3 Discrimitive

5.4 Evaluierung

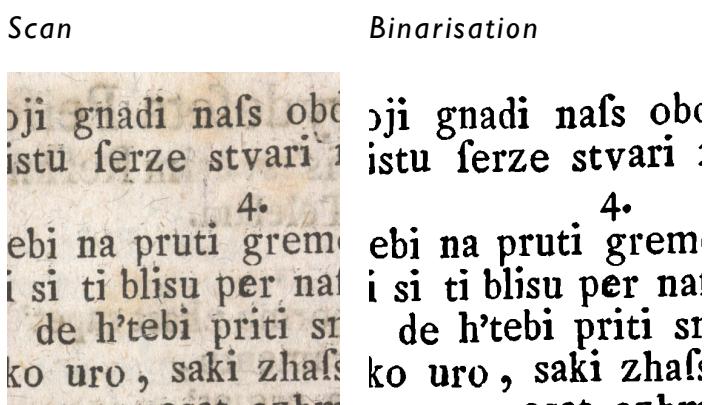


Abbildung 5.3: Beispiel aus dem DIBCO2013-Dateset

Metriken

Die Evaluierung der Ergebnisse der Segmentierung erfolgt auf Pixelbene. Long, Shelhamer und Darrell 2015 berechnet 4 Metriken. Sei n_{ij} die Anzahl der Pixel der Klasse i die der Klasse j zugeordnet wurden.

5.5 Fazit

Literaturverzeichnis

- Achanta, Radhakrishna u. a. (2010). „SLIC Superpixels“. In: URL: <https://infoscience.epfl.ch/record/149300> (besucht am 01.12.2017).
- Achanta, R. u. a. (2012). „SLIC Superpixels Compared to State-of-the-Art Superpixel Methods“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11, S. 2274–2282. DOI: [10.1109/TPAMI.2012.120](https://doi.org/10.1109/TPAMI.2012.120).
- Alighieri, Dante (1300). *Cologny, Fondation Martin Bodmer, Cod. Bodmer 55 : Dante, Inferno e Purgatorio (Codex Guarneri)*. DOI: [10.5076/e-codices-cb-0055](https://doi.org/10.5076/e-codices-cb-0055).
- Ambrosius, Mediolanensis u. a. (0985). *St. Gallen, Stiftsbibliothek, Cod. Sang. 18 : Composite Manuscript, Astronomical Clock of Pacificus of Verona*. e-codices - Virtual Manuscript Library of Switzerland. DOI: [10.5076/e-codices-csg-0018](https://doi.org/10.5076/e-codices-csg-0018). URL: <https://www.e-codices.ch/en/list/one/csg/0018> (besucht am 13.03.2018).
- Baird, Henry S. (2003). „Digital Libraries and Document Image Analysis“. In: *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference On*. IEEE, S. 2–14.
- Baird, Henry S. und Karl Tombre (2014). „The Evolution of Document Image Analysis“. In: *Handbook of Document Image Processing and Recognition*. Hrsg. von David Doermann und Karl Tombre. London: Springer London, S. 63–71. DOI: [10.1007/978-0-85729-859-1_43](https://doi.org/10.1007/978-0-85729-859-1_43). URL: http://link.springer.com/10.1007/978-0-85729-859-1_43.
- Baird, Professor Henry S. (2014). „A Brief History of Documents and Writing Systems“. In: *Handbook of Document Image Processing and Recognition*. Hrsg. von David Doermann und Karl Tombre. London: Springer London, S. 3–10. DOI: [10.1007/978-0-85729-859-1_2](https://doi.org/10.1007/978-0-85729-859-1_2). URL: http://link.springer.com/10.1007/978-0-85729-859-1_2 (besucht am 06.04.2018).
- Chen, Kai und Mathias Seuret (2017). „Convolutional Neural Networks for Page Segmentation of Historical Document Images“. In: arXiv: [1704.01474 \[cs, stat\]](https://arxiv.org/abs/1704.01474). URL: <http://arxiv.org/abs/1704.01474>.
- Chen, Kai, Mathias Seuret u. a. (2017). „Convolutional Neural Networks for Page Segmentation of Historical Document Images“. In: IEEE, S. 965–970. DOI: [10.1109/ICDAR.2017.161](https://doi.org/10.1109/ICDAR.2017.161).

- Chen, K. u. a. (2016). „Page Segmentation for Historical Document Images Based on Superpixel Classification with Unsupervised Feature Learning“. In: *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. 2016 12th IAPR Workshop on Document Analysis Systems (DAS), S. 299–304. DOI: [10.1109/DAS.2016.13](https://doi.org/10.1109/DAS.2016.13).
- Chollet, François (2018). *Deep Learning with Python*. OCLC: 982650571. Shelter Island, NY: Manning Publications Co.
- David S. Doermann, Hrsg. (2014). *Handbook of Document Image Processing and Recognition*. London: Springer Reference.
- Davidge (2018). *Cuneiform Script2.Jpg - Wikimedia Commons*. URL: https://commons.wikimedia.org/wiki/File:Cuneiform_script2.jpg#filelinks (besucht am 06.04.2018).
- Doersch, Carl und Andrew Zisserman (2017). „Multi-Task Self-Supervised Visual Learning“. In: arXiv: [1708.07860 \[cs\]](https://arxiv.org/abs/1708.07860). URL: <http://arxiv.org/abs/1708.07860>.
- Dosovitskiy, Alexey u. a. (2016). „Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.9, S. 1734–1747. DOI: [10.1109/TPAMI.2015.2496141](https://doi.org/10.1109/TPAMI.2015.2496141).
- e-codices (2018). *E-Codices – Virtuelle Handschriftenbibliothek Der Schweiz*. URL: <https://www.e-codices.unifr.ch/de/about/history> (besucht am 25.04.2018).
- Goodfellow, Ian, Yoshua Bengio und Aaron Courville (2016). *Deep Learning*. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press. 775 S.
- IIIF (2018). *International Image Interoperability Framework*. URL: <http://iiif.io/> (besucht am 29.03.2018).
- Krizhevsky, Alex, Ilya Sutskever und Geoffrey E Hinton (2012). „ImageNet Classification with Deep Convolutional Neural Networks“. In: *Advances in Neural Information Processing Systems 25*. Hrsg. von F. Pereira u. a. Curran Associates, Inc., S. 1097–1105. URL: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.
- LeCun, Yann, Yoshua Bengio und Geoffrey Hinton (2015). „Deep Learning“. In: *Nature* 521.7553, S. 436–444. DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539). pmid: [26017442](https://pubmed.ncbi.nlm.nih.gov/26017442/).
- LeCun, Yann, Bernhard Boser u. a. (1989). „Backpropagation Applied to Handwritten Zip Code Recognition“. In: *Neural computation* 1.4, S. 541–551.
- Long, Jonathan, Evan Shelhamer und Trevor Darrell (2015). „Fully Convolutional Networks for Semantic Segmentation“. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, S. 3431–3440.
- Lucanus, Marcus Annaeus (1025). *St. Gallen, Stiftsbibliothek, Cod. Sang. 863 : Pharsalia Libri Decem*. e-codices - Virtual Manuscript Library of Switzerland. DOI: [10.5076/e-codices-csg-0863](https://doi.org/10.5076/e-codices-csg-0863). URL: <https://www.e-codices.ch/en/list/one/csg/0863> (besucht am 13.03.2018).

- McConaughay, Lara, Jennifer Dai und David Bamman (2017). „The Labeled Segmentation of Printed Books“. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, S. 748–758.
- Mitchell, Tom M. (1997). *Machine Learning*. International ed., [Reprint.] McGraw-Hill series in computer science. OCLC: 846511832. New York, NY: McGraw-Hill. 414 S.
- Noroozi, Mehdi und Paolo Favaro (2016). „Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles“. In: arXiv: 1603.09246 [cs]. URL: <http://arxiv.org/abs/1603.09246>.
- Opitz, Andrea und Thomas Stäcker (2009). „Workshop Der Massendigitalisierungsprojekte Der Deutschen Forschungsgemeinschaft an Der Herzog August Bibliothek Wolfenbüttel“. In: *Zeitschrift für Bibliothekswesen und Bibliographie* 56.6, S. 363–373. DOI: 10.3196/186429500956641.
- Paszke, Adam u. a. (2017). „Automatic Differentiation in PyTorch“. In: URL: <https://openreview.net/forum?id=BJJsrmfCZ> (besucht am 11.03.2018).
- Rawat, Waseem und Zenghui Wang (2017). „Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review“. In: *Neural Computation* 29.9, S. 2352–2449. DOI: 10.1162/neco_a_00990.
- Simistira, Foteini u. a. (2016). „DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts“. In: IEEE, S. 471–476. DOI: 10.1109/ICFHR.2016.0093.
- Simistira, F. u. a. (2017). „ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts“. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Bd. 01, S. 1361–1370. DOI: 10.1109/ICDAR.2017.223.
- Smith, PhDElisa H. Barney (2014). „Document Creation, Image Acquisition and Document Quality“. In: *Handbook of Document Image Processing and Recognition*. Hrsg. von David Doermann und Karl Tombre. London: Springer London, S. 11–61. DOI: 10.1007/978-0-85729-859-1_3. URL: http://link.springer.com/10.1007/978-0-85729-859-1_3 (besucht am 06.04.2018).
- Staatsbibliothek zu Berlin (2016). *Digitalisierte Sammlungen Der Staatsbibliothek Zu Berlin*. URL: <http://digital.staatsbibliothek-berlin.de/> (besucht am 21.07.2016).
- Sun, Chen u. a. (2017). „Revisiting Unreasonable Effectiveness of Data in Deep Learning Era“. In: arXiv: 1707.02968 [cs]. URL: <http://arxiv.org/abs/1707.02968>.
- Süße, Herbert und Erik Rodner (2014). *Bildverarbeitung und Objekterkennung: Computer Vision in Industrie und Medizin*. OCLC: 897913586. Wiesbaden: Springer Vieweg. 666 S.
- Valveny, Ernest (2014). „Datasets and Annotations for Document Analysis and Recognition“. In: *Handbook of Document Image Processing and Recognition*. Hrsg. von David Doermann und Karl

- Tombre. London: Springer London, S. 983–1009. DOI: [10.1007/978-0-85729-859-1_32](https://doi.org/10.1007/978-0-85729-859-1_32). URL: http://link.springer.com/10.1007/978-0-85729-859-1_32 (besucht am 06.03.2018).
- Wang, Xiaolong und Abhinav Gupta (2015). „Unsupervised Learning of Visual Representations Using Videos“. In: *The IEEE International Conference on Computer Vision (ICCV)*.
- Wick, Christoph und Frank Puppe (2017). „Fully Convolutional Neural Networks for Page Segmentation of Historical Document Images“. In: arXiv: [1711.07695 \[cs\]](https://arxiv.org/abs/1711.07695). URL: <http://arxiv.org/abs/1711.07695>.
- Xu, Y. u. a. (2017). „Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks“. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Bd. 01, S. 541–546. DOI: [10.1109/ICDAR.2017.94](https://doi.org/10.1109/ICDAR.2017.94).

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle aus fremden Quellen im Wortlaut oder dem Sinn nach entnommenen Aussagen sind durch Angaben der Herkunft kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Ort, Datum

Unterzeichner