

Multimodaler Bedeutungstransfer vom Text zum Bild. Granulare Bildklassifikation durch Verteilungssemantik.

Donig, Simon

simon.donig@uni-passau.de

Universität Passau, Deutschland

Maria, Christoforaki

christoforakimaria@gmail.com

Universität Passau, Deutschland

Bernhard, Bermeitinger

bernhard.bermeitinger@unisg.ch

Universität Sankt Gallen, CH; Universität Passau, Deutschland

Handschuh, Siegfried

siegfried.handschuh@unisg.ch

Universität Sankt Gallen, CH; Universität Passau, Deutschland

Einleitend

In den letzten Jahren hat die Verwendung von Bildklassifizierungverfahren wie Neuronalen Netzen auch im Bereich der historischen Bildwissenschaften und der Heritage Informatics weite Verbreitung gefunden (Lang, Ommer 2018). Diese Verfahren stehen dabei vor einer Reihe von Herausforderungen, darunter dem Umgang mit den vergleichsweise kleinen Datenmengen sowie zugleich hochdimensionalen Datenräumen in den digitalen Geisteswissenschaften. Meist bilden diese Methoden die Klassifizierung auf einen vergleichsweise flachen Raum ab. Dieser „flache“ Zugang verliert im Bemühen um ontologische Eindeutigkeit eine Reihe von relevanten Dimensionen, darunter taxonomische, mereologische und assoziative Beziehungen zwischen den Klassen beziehungsweise dem nicht formalisierten Kontext. Eine in (Donig, Christoforaki, Bermeitinger, Handschuh 2019) vorgeschlagene Lösung, diese Beziehungen wieder in den Prozess der Klassifizierung zurückzubringen, ist, sich die größere Ausdruckskraft von textbasierten Modellen zunutze zu machen, um die Fähigkeiten visueller Klassifikatoren zu erweitern.

Dabei wird ein Convolutional Neural Network genutzt, dessen Ausgabe im Trainingsprozess anders als

herkömmlich nicht auf einer Serie flacher Textlabel beruht, sondern auf einer Serie von Vektoren. Diese Vektoren resultieren aus einem Distributional Semantic Model (DSM), welches aus einem Domäne-Textkorpus generiert wird. Ein DSM ist ein multidimensionaler Vektorraum, in dem Wörter als Vektoren abgebildet werden (Lenci 2018). Wir stellen hier eine frühe Implementierung des Verfahrens vor und analysieren deren Ergebnisse.

Wir stellen hier eine frühe Implementierung des Verfahrens vor und analysieren deren Ergebnisse.

Das durchgeführte Experiment beruht auf der Kollation von zwei Korpora, einem textbasierten und einem visuellen. Mit dem Textkorpus wird zunächst ein DSM erzeugt und diesem dann eine Auswahlliste von Zielwörter zugeführt (die funktional den Annotationslabeln der Bilder entspricht). Als Ergebnis erhalten wir Vektoren, die mit diesen Wörtern korrespondieren und mit denen die Bilder annotiert werden. Mit diesen Vektorannotationen wird ein neuronales Netzwerk trainiert, das anschließend dem Klassifikator unbekanntes Bildmaterial identifizieren soll. Als Ergebnis dieses Klassifikationsprozesses erhalten wir einen Vektor, der mit Hilfe des DSMs in natürlichsprachige Wörter zurückgewandelt wird. Da wir nach reichereren Repräsentationen im Zuge dieses Vorgangs suchen, wählen wir die fünf Vektoren aus, die dem Ausgangsvektor am ähnlichsten sind (Top-5 Nearest Neighbours). Als Ähnlichkeitsmaß legen wir die Kosinus-Ähnlichkeit zwischen vorhergesagtem Vektor und jenem Vektor zugrunde, der dem ursprünglich dem Bild von uns zugewiesenen Label (Goldlabel) entspricht. Wir gehen davon aus, dass ein Bild korrekt klassifiziert wurde, wenn das Goldlabel unter den Top-5 erscheint.

Darüber hinaus vergleichen wir die Ergebnisse des vorgeschlagenen Klassifizierungsverfahrens mit einem herkömmlichen Verfahren auf der Grundlage flacher Label unter Verwendung desselben CNNs, das für das Vektor-Experiment genutzt wurde. Wir können zeigen, dass das Vektor-Verfahren (bezogen auf die Metriken) ebenso effizient und in einigen Aspekten sogar besser ist.

Aufbau des Experiments

Das Experiment beruht auf je einem Bild- und Textkorpus aus dem Bereich Sachkulturforschung mit einem Fokus auf klassizistische Artefakte.

Das Textkorpus besteht aus 44 Quellen, die unter einer freien, permissiven Lizenz verfügbar sind, und umfasst englischsprachige Fachpublikationen zu Mobiliar und Raumkunst, die von der Jahrhundertwende bis zur Mitte des 20. Jahrhunderts erschienen sind. Das Textkorpus wurde in mehreren Schritten gereinigt und aufbereitet. Zum einen wurden Standard-Natural-Language-Processing-Verfahren (NLP) angewandt, darunter Tokenisierung, Satz- und Worttrennung, die Normalisierung von Zahlenwerten und die Erkennung von benannten Entitäten (NER). Da wir retrodigitalisiertes Material aus verschiedenen Quellen genutzt haben, implementierten wir manuelle

Korrekturen für die häufigsten der vorkommenden Fehler (etwa Ligaturen wie II, die als U fehlinterpretiert wurden). Eine weitere Ebene der Vorverarbeitung bestand aus inhaltsbezogenen Augmentierungen. Insbesondere normalisierten wir zusammengesetzte Wörter und Synonyme gemäß einer spezifizierten Liste, die anhand einer Ontologie, der Neoclassica-Ontologie (Donig, Christoforaki, Handschuh 2016) zusammengestellt wurde. Dies resultierte in einem Korpus von 3.067.237 Wörtern aus 107.518 Wortgrundformen.

Das DSM wurde von uns mit Hilfe des Indra Frameworks (Sales, Souza, Barzegar, Davis, Freitas, Handschuh 2018) und Gensim (#eh##ek und Sojka 2010) erzeugt.¹

Das Bildkorpus besteht aus 1231 Ansichten klassizistischer Möbel in deren Gesamtheit, die permissiv lizenziert sind² und die sowohl historisches Bildmaterial als auch Fotos aus der modernen Bestandsdokumentation umfassen. Es repräsentiert 28 Klassen.

Da es sich um ein *Proof-of-Concept* Experiment handelt, kam zum Zweck des Rapid Prototyping ein an die VGG-Architektur angelehntes, „simples“ neuronales Netzwerk zum Einsatz.³ Die Unabhängigkeit der Trainings- und Testbeispiele wurde durch einen Train/Test/Eval-Split von 55:20:25 garantiert.

Da durch Sammlungspraxis der Gedächtnisinstitutionen (Sammelwürdigkeit, geographischer Schwerpunkt) und Zugänglichkeit des Materials (Lizenzierung, Grad der Sammlungsdigitalisierung) die Verteilung der Artefakte nach Klassen unbalanciert ist (Abb. 1), haben wir die Klassengewichte dementsprechend angepasst (seltene Klassen werden höher gewichtet als häufig vorkommende Klassen (Johnson, Khoshgoftaar 2019: 27). Um eine Situation zu vermeiden, in der ein Machine Learning Modell derart an ein Eingabedaten-Set angepasst wird, dass es darin scheitert, auf ähnlichen Daten zu generalisieren (Overfitting), wurde die übliche Early-Stopping-Methode verwendet.

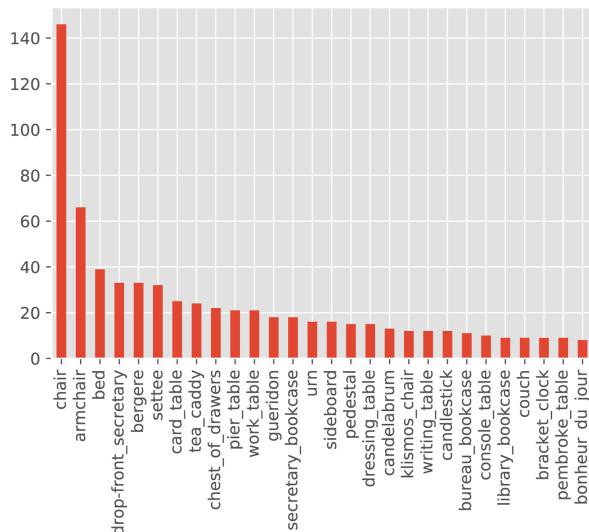


Abbildung 1: Verteilung der Klassen im Bildkorpus

Ergebnisse

Die Top-5-Richtig-Positiv Rate betrug 0.59. Das bedeutet, dass das Goldlabel in 59% der Fälle unter den fünf nächsten Nachbarn erschien.

Das mathematische Qualitätskriterium gibt für sich genommen jedoch nur einen Teil des Gesamtbilds wieder. Wir haben deshalb zugleich eine qualitative Analyse der Ergebnisse im Testset durchgeführt.

Eine Reihe von richtig-positiven Ergebnissen zeigen etwa, dass die Klassifizierung keinesfalls zufällig erfolgt, sondern dass die Top-5-Begriffe tatsächlich jeweils denselben semantischen Nachbarschaften entstammen. Sie drücken eine Reihe von Beziehungen taxonomischer und assoziativer Natur aus.

Beispielsweise wird der Roentgen-Schreibtisch aus dem Bestand des V&A in Abb. 2 mit Labeln (in der Reihenfolge) von dressing_table, writing_table und work_table assoziiert (Ankleidetisch, Schreibtisch, Nähtisch). Diese Trias ist schon deshalb sinnhaftig, weil viele dieser Artefakte multifunktional waren und mehrere dieser Funktionen erfüllten. Daneben ähneln auch jene Artefakte, die deziidiert nur einem einzigen Zweck dienten, konstruktiv den jeweils anderen Möbeltypen. Die Nähe der drei Konzepte entsteht also sowohl auf semantischer Ebene (Nähe der Wörter im DSM, die wiederum das Produkt lebensweltlicher Nähe ist), als auch auf einer visuellen Ebene im CNN (visuelle Formähnlichkeit). Ein weiteres Bild desselben Objekts (Abb. 3) zeigt einerseits, dass die Methode in sich konsistent ist (die Top4 sind identisch, obwohl eine andere Perspektive vorliegt) und andererseits, dass auch die visuellen Merkmale innerhalb des CNNs eine Auswirkung auf den Klassifizierungsprozess haben. Da Schreibschränke (*secrétaire à abbatants*) häufig frontal, hochaufricht und mit einer geöffneten Schreibklappe oder -schublade abgebildet werden, scheint deren Vorkommen im Bild eine Klassifizierung als Sekretär getriggert zu haben. Im ersten Bild könnte dagegen die Anwesenheit von Schubladen (*drawers*) zu einer Klassifizierung als Kommode geführt haben, die naheliegenderweise auf semantischer Ebene mit Schubladen assoziiert ist.



Abbildung 2: Abweichungen bei der Klassifizierung desselben Objekts



Abbildung 3: Abweichungen bei der Klassifizierung desselben Objekts

Während die Label in den bisher betrachteten Fällen die taxonomischen Beziehungen reflektieren und alle den aus der Ontologie abgeleiteten Target Words entstammen, zeigt Abb. 4, dass das Verfahren auch aus sich selbst, rein datenzentriert Label generieren kann. Die abgebildete Kratervase wurde als Urne (urn) gold-klassifiziert. Die Top-2 Wörter reflektieren demnach auch taxonomischen Beziehungen (urn, vase). Die anderen Konzepte spiegeln dagegen assoziative Beziehungen wider. Das Label „bell“ ist ein Artefakt des Reinigungsprozesses, da im Korpus Wörter wie „bell-shaped, bell-crater“ (mit und ohne Bindestrich) existieren, um diese Art von Artefakten zu beschreiben. „Ovoid“ bezieht sich demgegenüber wohl auf die Eierstabdekoration des oberen Wulsts, die oft mit diesem Adjektiv beschrieben wird. Diese Ornamentik scheint zugleich die Assoziation zur Rosette (Patera) mitbedingt zu haben. Auf diese Weise erscheint das Target Word „patera_element“ unter den Top-5, obwohl im Bildkorpus ausschließlich ganze Artefakte, nicht aber deren Dekor annotiert wurden.



Abbildung 4: Eine Sèvres-Kopie der Medici-Vase stößt die Klassifizierung mit assoziativen Labeln an

Nicht auszuschließen ist hier zudem ein Effekt des visuellen Klassifikators, wie auch Abb. 5 zeigt. Die Fehlklassifizierung des Objekts, eines Nähtischchens, führte zu konsistenten Zuschreibungen im Bereich der Sitz- und Liegemöbel. Betrachtet man die äußere Form des Artefakts auf einer abstrakteren Ebene, kann man eine visuelle Nähe zu z.B. einem (Double-)Camel-back Sofa durchaus nachvollziehen.

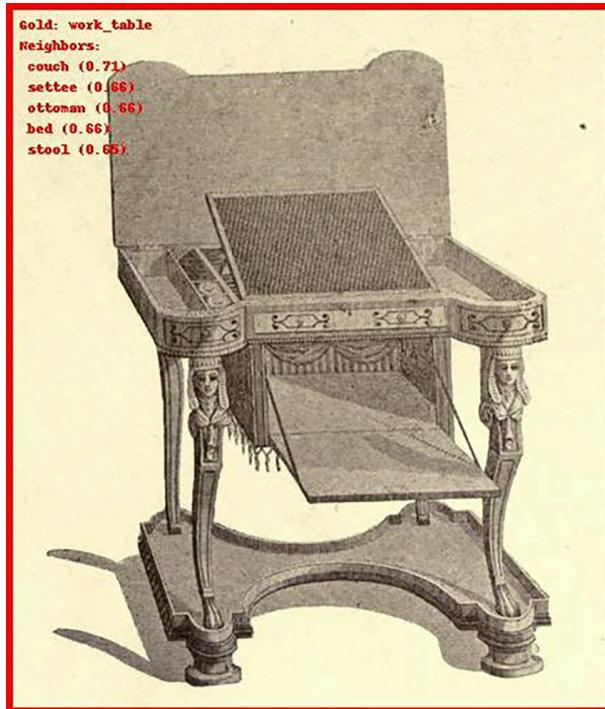


Abbildung 5: Fehlklassifizierung eines Nähtischs in ein Wortumfeld aus der Sitzmöbel-Hierarchie

Vergleichsexperiment mit einem CNN mit flachen Labeln

Um die Unterschiede zwischen beiden Zugängen besser abschätzen zu können, haben wir weiter ein Vergleichsexperiment durchgeführt, bei dem wir dasselbe CNN wie im verkörperten Verfahren für eine herkömmliche Klassifizierung mit flachen Labeln heranzogen.⁴

Tabelle 1: Vergleich zentraler Metriken beider Zugänge

	Vektor-Label	Flache Label
Top-1-Treffergenauigkeit	0.50	0.40
Top-1-Genauigkeit	0.32	0.29
Top-1-Trefferquote	0.25	0.26
Top-1-F1-Wert	0.26	0.25
Top5-Falsch-Positiv-Rate	0.41	0.27
Top5-Richtig-Positiv-Rate	0.59	0.73

Wie ersichtlich, ist nicht nur die Top-1-Treffergenauigkeit im Fall der Klassifizierung mit

Vektoren besser, sondern auch die übrigen Metriken vergleichbar gut sind. Durch den hier vorgeschlagenen Zugang wird also nicht nur die Treffergenauigkeit verbessert, sondern er liefert zugleich eine reichhaltigere Beschreibung des Bildes.

Schlussfolgerungen und Ausblick

In dem hier vorgeschlagenen Beitrag haben wir ein neues, multimodales Verfahren für die Klassifizierung von Bildinhalten vorgestellt, das auf der Kombination von NLP-Methoden mit Bildklassifizierungsverfahren beruht. Ziel war, Objekte nicht alleine nach einem Schema flacher Label, sondern in einer kontextgerechteren Weise zu klassifizieren, wobei dieser Kontext von einschlägigen historischen Domänenpublikationen gebildet wird. Dieses Klassifizierungsverfahren bietet einen Zugang zur multidimensionalen Einbettung der Artefakte in die Lebenswelt und deren sprachlicher Widerspiegelung. Dieser Umstand ist von besonderem Nutzen, um multifunktionale Objekte zu klassifizieren, ohne dabei auf mehrere Klassifikatoren und einen komplexen Annotationsprozess mit mehreren Labeln zurückgreifen zu müssen. Die Ergebnisse sind ermutigend. Auch mit einem sehr einfachen CNN erreichten wir eine Genauigkeit von 0,59. Als nächsten Schritt möchten wir mit einem komplexeren CNN und einem ausgeweiteten Bildkorpus trainieren (um bekannte Probleme wie overfitting zu reduzieren). Unser Vergleichsexperiment mit einem herkömmlichen, auf flachen Labeln beruhenden Zugang hat gezeigt, dass unter Effizienzgesichtspunkten, d. h. im direkten Vergleich der Metriken, unser Verfahren nicht nur vergleichbare Resultate liefert, sondern zugleich auch in einer reichhaltigeren Beschreibung des Bildes resultiert.

Wir werden weiter daran arbeiten, besser zu verstehen, wie ein bestimmtes Textkorpus sich in den Labeln widerspiegelt, die das DSM automatisch zuweist und die nicht Teil der Liste der Target Words sind. Ein besseres Verständnis dieser Prozesse scheint insbesondere im Hinblick auf die relativ überschaubaren Textkorpora relevant, die in den Geisteswissenschaften zu spezifischen Themenkomplexen kollationiert werden können. Nicht zuletzt werden wir aus diesem Grund die Nutzung von Thesauri und Wörterbüchern in Betracht ziehen, um Synonymlisten für Target Words zu erstellen. In ähnlicher Weise ziehen wir in Betracht, benannte Entitäten zu URIs zusammenzufassen. Das würde uns erlauben, spezifische Entitäten (z. B. Werkstätten, Ebenisten, Eigentümer) mit bestimmten Objekten zu assoziieren.

Wir denken, dass dabei der multimodale Zugriff einen besonders effizienten Zugang zu geistes- und kulturwissenschaftlichen Korpora bietet, die, verglichen mit den Korpora anderer Disziplinen in den Natur- und Sozialwissenschaften, klein und Domäne-restringiert sind.

Fußnoten

1. Das DSM wurde mit einer Vektorgröße von 50, einer Wortfenstergröße von 10 und einer minimalen Wortzahl von fünf erstellt. Als Word2Vec-Modell kam Skipgram (Mikolov, Chen, Corrado, Dean 2013) mit Negative Sampling zum Einsatz.
2. Das Korpus wurde aus den Sammlungen des Metropolitan Museum, New York, des Victoria & Albert Museum, London, der Wallace Collection, London sowie mehreren zeitgenössischen Musterbüchern zusammengestellt.
3. Das Netzwerk besteht aus drei Convolutional-Blöcken mit jeweils zwei Convolutional-Layers mit 32/64/64 Filter der Größe 3x3. Nach jedem Block folgt ein Maximum-Pooling-Layer der Größe 2x2 sowie ein Dropout-Layer mit einer Dropoutwahrscheinlichkeit von 0.25. Ein Fully-Connected-Block, bestehend aus zwei Fully-Connected-Layers mit jeweils 256 Knoten, steht im Anschluss sowie nochmals ein Dropout Layer mit 0.5 Dropoutwahrscheinlichkeit. Jeder Convolutional- und Fully-Connected-Layer bis dahin wurde zufällig initialisiert und benutzt ReLU als Aktivierungsfunktion. Der letzte Layer ist ein Fully-Connected-Layer mit 50 Ausgabeknoten und benutzt eine lineare Aktivierungsfunktion. Es ist in den beiden von uns genutzten Frameworks Keras (<https://keras.io>) und TensorFlow (<https://tensorflow.org>) implementiert. Beim Training wird der durchschnittliche absolute Fehler durch die Optimierungsfunktion RMSprop minimiert.
4. Das verwendete CNN unterscheidet sich von dem im Vektor-Experiment verwendeten Netz lediglich durch den letzte Layer, der ein Fully-Connected-Layer mit 28 Ausgabeknoten ist, korrespondierend mit den 28 flachen Labeln, und die Nutzung von softmax als Aktivierungsfunktion.

Bibliographie

Donig, Simon / Christoforaki, Maria / Handschuh, Siegfried (2016): “Neoclassica – A Multilingual Domain Ontology. Representing Material Culture from the Era of Classicism in the Semantic Web”, in: Bozic, Bojan/Mendel-Gleason, Gavin/Debruyne, Christophe / O’Sullivan, Declan (eds.): Computational History and Data-Driven Humanities. CHDDH 2016 (=IFIP Advances in Information and Communication Technology, vol 482), Cham: Springer: 41–53, DOI 10.1007/978-3-319-46224-0_5 [Letzter Zugriff 25. 09. 2019].

Donig, Simon / Christoforaki, Maria / Bermeitinger, Bernhard / Handschuh, Siegfried (2019): „Vom Bild zum Text und wieder zurück“, in: Sahle, Patrick (ed.): DHd 2019 - Digital Humanities: multimedial & multimodal – Konferenzabstracts. Mainz & Frankfurt a. M.: 227–232, <https://zenodo.org/record/2596095/>

files/2019_DHd_BookOfAbstracts_web.pdf [Letzter Zugriff 25. 09. 2019].

Johnson, Justin M. / Khoshgoftaar, Taghi M. (2019): „Survey of deep learning with class imbalance“, in: Journal of Big Data 6 (27): 2-54, <https://doi.org/10.1186/s40537-019-0192-5> [Letzter Zugriff 25. 09. 2019].

Krizhevsky, Alex / Sutskever, Ilya / Hinton Geoffrey E. (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. <http://dl.acm.org/citation.cfm?id=2999134.2999257> [Letzter Zugriff 25. 09. 2019].

Lang, Sabine / Ommer, Björn (2018): „Attesting Similarity: Supporting the Organization and Study of Art Image Collections with Computer Vision“. In: Digital Scholarship in the Humanities 33 (4): 845–56. <https://doi.org/10.1093/llc/fqy006> [Letzter Zugriff 25. 09. 2019].

Lenci, Alessandro (2018): “Distributional models of word meaning”, in: Annual review of Linguistics, 4 (1) : 151-171.

Mikolov, Tomas / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): “Efficient Estimation of Word Representations in Vector Space.” ArXiv:1301.3781 [Cs]. <http://arxiv.org/abs/1301.3781>. [Letzter Zugriff 25. 09. 2019]

#eh##ek, Radim / Sojka, Petr (2010): „Software Framework for Topic Modelling with Large Corpora“. In: LREC 2010. Valletta, Malta, 2010: 46–50.

Sales, Juliano Efson / Souza, Leonardo / Barzegar, Siamak / Davis, Brian / Freitas, André / Handschuh, Siegfried (2018) “Indra: A Word Embedding and Semantic Relatedness Server.” In LREC. Miyazaki, Japan, 2018.

Abb. 2 / Abb. 3: Victoria & Albert Museum, London: Writing table, Neuwied, workshop of David Roentgen ca. 1774-1780. Ascension number: 1059:1 to 9-1882. <http://collections.vam.ac.uk/item/O117298/writing-table-roentgen-david/> [Letzter Zugriff 25. 09. 2019].

Abb. 4: Victoria & Albert Museum, London: Vase [Sèvres Copy of the Medici Vase], Paris, 1813. Ascension Number 396-1874. <http://collections.vam.ac.uk/item/O8978/vase-sevres-porcelain-factory/> [Letzter Zugriff 25. 09. 2019].

Abb. 5: **Sheraton, Thomas / Bell, J. Munro (arr.)** (1910): The furniture designs of Thomas Sheraton. London: Gibbings and Co., Ltd.