
MULTIMODAL SEMANTIC TRANSFER FROM TEXT TO IMAGE. FINE-GRAINED IMAGE CLASSIFICATION BY DISTRIBUTIONAL SEMANTICS.

Simon Donig 
 Chair for Digital Humanities
 University Passau, Germany
 simon.donig@uni-passau.de

Maria Christoforaki
 Chair for Data Science
 Institute for Computer Science
 University of St.Gallen, Switzerland
 maria.christoforaki@unisg.ch

Bernhard Bermeitinger 
 Chair for Data Science
 Institute for Computer Science
 University of St.Gallen, Switzerland
 bernhard.bermeitinger@unisg.ch

Siegfried Handschuh
 Chair for Data Science
 Institute for Computer Science
 University of St.Gallen, Switzerland
 siegfried.handschuh@unisg.ch

December 2019

1 Introduction

In the last years, image classification processes like neural networks in the area of art-history and *Heritage Informatics* have experienced a broad distribution (Lang and Ommer 2018). These methods face several challenges, including the handling of comparatively small amounts of data as well as high-dimensional data in the Digital Humanities. In most cases, these methods map the classification task to flat target space. This “flat” surface loses several relevant dimensions in the search for ontological uniqueness, including taxonomical, mereological, and associative relationships between classes, or the non-formal context, respectively.

The proposed solution by Donig, Bermeitinger, et al. (Donig, Bermeitinger, et al. 2019) to expand the capabilities of visual classifiers is to take advantage of the greater expressiveness of text-based models. Here, a *Convolutional Neural Network* (CNN) is used that output is not as usual a series of flat text labels but a series of semantically loaded vectors. These vectors result from a *Distributional Semantic Model* (DSM) (Lenci 2018) which is generated from an in-domain text corpus.

Here, we propose an early implementation of this method and analyze the results.

The conducted experiment is based on the collation of two corpora: one text-based and a visual. From the text, a DSM is created and then queried for a list of target words that are functionally the labels that are manually given to the images. The result is a list of vectors that correspond to the target words leading to images that are annotated not only with a label but also with a unique vector. The images and vectors are used as training data for a CNN that, afterward, should be able to predict a vector for an unseen image. This prediction vector can be converted back to a word by the DSM using a nearest-neighbor algorithm. We are looking for richer representation in this process, so we choose the five nearest neighbors. The similarity measure is the cosine similarity for high-dimensional vector spaces between the given target vector and the prediction vector. We derive a positive classification result if the target label is within the list of five nearest neighbors of the prediction vector.

Moreover, we compare the results between the proposed classification method and a conventional classification method using the same CNN as for the vector-based experiment but a list of flat labels. Finally, we can show that the vector-based approach (judging from classification metrics) is equally performant or even better than the label-based version.

2 Experiment Structure

The experiment is based on one text and one visual corpus from the area of material culture research with a focus on neo-classical artifacts.

2.1 Text Corpus

The text corpus is compiled out of 44 sources that are available under a free and permissive license. It contains English specialist publications on furniture and spatial art, published from the end of the 19th century to the middle of the 20th century. In multiple steps, the corpus is cleaned and preprocessed. First, a series of standard NLP methods are applied like tokenization, sentence- and word splitting, normalization of numbers, and named entity recognition (NER). Since we used retro-digitized material from a different source, we implemented manual corrections for the most common errors (such as ligatures like TT that were misinterpreted as U). Another level of preprocessing consists of content-related augmentations. In particular, we normalized compound words and synonyms according to a specified list, which is based on an ontology, the *Neoclassica-Ontology* (Donig, Christoforaki, and Handschuh 2016). This resulted in the final text corpus of total 3 067 237 words comprised of 107 518 basic word forms.

The DSM is created using the *Indra Frameworks* (Sales et al. 2018) with a vector size of 50, a word window size of 10, and minimal word count of 5. We used *Skipgram* as the *Word2Vec* model (Mikolov et al. 2013) with negative sampling.

2.2 Image Corpus

The image corpus consists of 1231 images of neoclassical furniture in their entirety, which are permissive licensed¹. The images are either historical pictorial material or

¹The corpus was compiled from collections from the Metropolitan Museum, New York, the Victoria & Albert Museum, London, the Wallace Collection, London, and several contemporary pattern books.

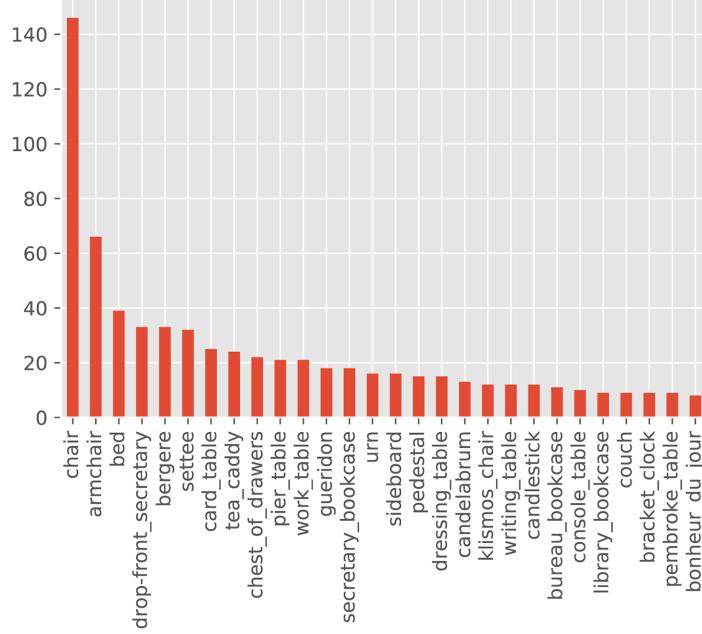


Figure 1: Class distribution of the image corpus

photographs from the modern as-built documentation. They are split into 28 different classes.

2.3 Combined Corpus

The nature of the experiment is *proof-of-concept*, so we used a VGG-like architecture of a “simple” convolutional neural network².

The independence and robustness of the train/test split are guaranteed with 5-fold cross-validation on 80 % on the full corpus from which 80 % are used for training and 20 % for testing. The remaining 20 % are treated as an evaluation set.

The dataset is unbalanced, as can be seen in Fig. 1. During training, more prominent classes are weighted down and underrepresented classes are given a higher weight (Johnson and Khoshgoftaar 2019, p. 27). Apart from dropout during training for regularization, *Early Stopping* was used to prevent overfitting.

²The CNN is built from three convolutional blocks with two consecutive Convolutional Layers each with 32/64/64 filters of size 3×3 . Each block is followed by a Maximum Pooling Layer with a size of 2×2 and a Dropout Layer for regularization with a dropout rate of 0.25. After the Convolutional Block, there are two Fully Connected Layers with 256 nodes and a Dropout Layer with a dropout rate of 0.5. The weights and biases for the Convolutional and the Fully Connected Layers are initialized randomly. Their activation function is *ReLU*. The last layer, the classification layer, is a Fully Connected Layer with 50 output nodes and a linear activation function. The optimization function of the chosen *mean absolute error* function is *RMSprop*. The implementation is done with Keras (<https://keras.io>) and TensorFlow (<https://tensorflow.org>)

3 Results

3.1 Results for CNN trained on Word Vectors

The Top-5 true-positive rate is 0.6052, meaning that the gold label from the annotations is in 60 % of the cases within the list of the five nearest neighbors.

However, the mathematical quality metric in itself represents only part of the overall picture. For this reason, a qualitative analysis of the results was performed.

A few true-positives show that, for example, the classification is by no means random but that the top 5 terms originate from the same semantic neighborhoods. They express several relationships of a taxonomical and associative nature.

As an example, the *Roentgen* desk from the *Victoria & Albert* inventory in Fig. 2 is associated with the labels `dressing_table`, `writing_table` and `work_table`. This triad is meaningful because many of these artifacts were multifunctional and fulfilled several of these functions. Besides, those artifacts that decidedly served only one purpose are constructively similar to the other types of furniture. The similarity of the three concepts thus emerges both on a semantic level (the distance of the words in the DSM, which in turn is the product of real-world distance) and on a visual level in the CNN (visual similarity of the form). Another image of the same object (see Fig. 3) shows, on the other hand, that the method is consisted in itself—the top 4 nearest neighbors of the predicted vector are identical, although the photograph is taken from a different perspective—and, on the other hand, that the visual features within the CNN also affect the classification process. Since writing cabinets (*secrétaires à abbatants*) are often displayed frontally, upright and with an open flap or drawer, their appearance in the image seems to have triggered a classification as a secretary. In the first image, however, the presence of drawers could have led to a classification as a chest-of-drawers, which is associated with drawers on a semantic level.

While the labels considered so far reflect the taxonomic relationships and all originate from the target words derived from the ontology, Fig. 4 shows that the procedure can also generate labels for itself, purely data-driven. The crater vase shown was originally classified as an urn. The Top-2 words therefore also reflect taxonomic relationships (“urn”, “vase”). The other concepts reflect associative relationships. The label `bell` is a leftover from the data cleaning process of the text corpus to describe this kind of artifact. “Ovoid” refers to the egg stick decoration of the upper bead, which is often described with this adjective. This ornamentation seems at the same time to have an association with the rosette (patera). In this way, the target word `patera_element` appears among the top-5 nearest neighbors, although only whole artifacts were annotated in the image corpus but not their decoration.

An effect of the visual classifier cannot be excluded, as shown in Fig. 5. The misclassification of the object, a small sewing table, led to consistent attributions in the area of seating and reclining furniture. Looking at the outer form of the artifact on a more abstract level, visual proximity to e.g. a (double)-camel-back sofa can be easily derived.



Figure 2: Differences in the classification of the same object (1)



Figure 3: Differences in the classification of the same object (2)



Figure 4: A sèvres copy of a Medici vase results in a classification of associative labels.

3.2 Comparison to a CNN with flat labels

To better estimate the differences between the two approaches (conventional classification by flat labels vs. vector-based approach), we present a comparison between different metrics for both approaches.

As shown in Table 1 the metrics are comparable.

Thus, the proposed approach not only improves the accuracy but also provides a richer description of the image.

4 Conclusion and Outlook

In this paper, we have presented a new multimodal approach for the classification of images based on the combination of NLP methods with image classification techniques. The goal was to classify objects not only according to a scheme of flat labels but in a more context-appropriate way, whereby the context informed by relevant domain-

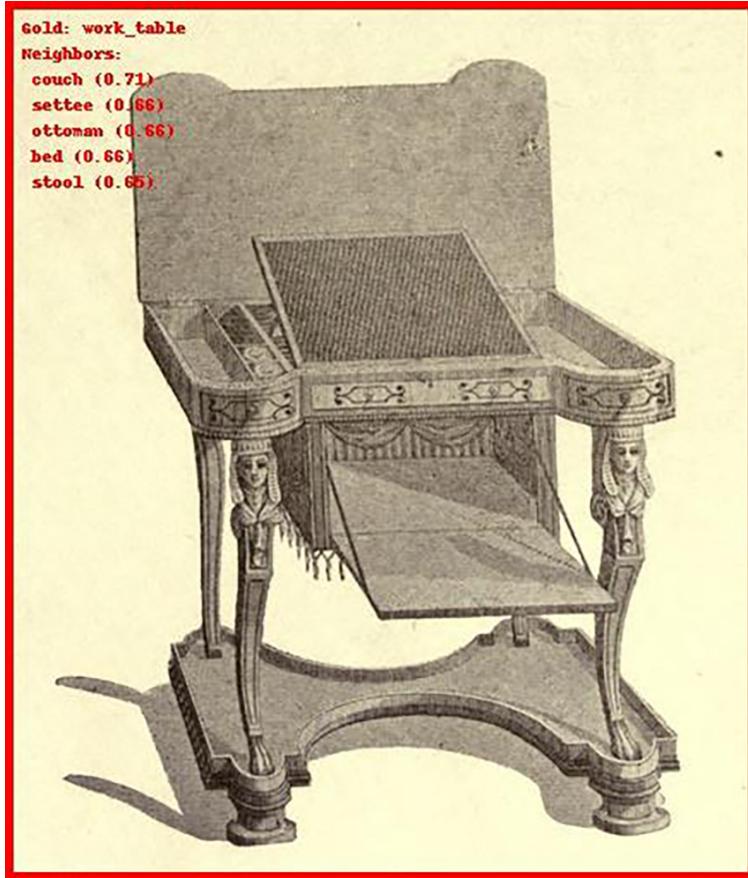


Figure 5: Misclassification of a sewing table in the proximity of seating furniture hierarchy.

Metric	Vector-based	Label-based
Top-1 Accuracy	0.5378	0.5302
Top-1 Precision	0.3487	0.4446
Top-1 Recall	0.2974	0.4044
Top-1 F1-Score	0.2976	0.3985
Top-5 True-Positive-Rate	0.6052	0.8463
Top-5 False-Positive-Rate	0.3948	0.1537

Table 1: Results of different Metrics.

historical publications. This classification method offers access to the multidimensional embedding of artifacts in the real-world and their linguistic reflection. This circumstance is particularly useful for classifying multifunctional objects without having to resort to several classifiers and a complex annotation process with several labels. The results are encouraging. Even with a very simple CNN, we achieve an accuracy of 0.54.

As a next step, we want to train with a deeper CNN and an extended image corpus—to reduce known problems like overfitting. The comparative experiment with a conventional flat-label approach has shown that from an efficiency point-of-view, i.e., indirect metric comparison, our method not only provides comparable results but also provides a richer description of the image.

We will continue to work to better understand how a particular body of text is reflected in the labels that the DSM automatically assigns and that are not part of the list of target words. A better understanding of these processes seems particularly relevant given the relatively manageable text corpora that can be collated into specific topic complexes in the humanities. Last but not least, for this reason, we will consider the use of thesauri and dictionaries to create synonym lists for target words. Similarly, we are considering combining named entities into URIs. This would allow us to associate specific entities (e.g., workshops, ebenists, owners) with specific objects.

We think that multimodal access provides particularly efficient access to humanities and cultural studies corpora that are small and domain-restricted compared to corpora of other disciplines in the natural and social sciences.

References

- Donig, Simon, Bernhard Bermeitinger, et al. (2019). “Vom Bild Zum Text Und Wieder Zurück”. In: *Digital Humanities: Multimedial & Multimodal - Konferenzabstracts*. DHd2019. Frankfurt am Main, Germany. URL: <https://www.researchgate.net/publication/332275547>.
- Donig, Simon, Maria Christoforaki, and Siegfried Handschuh (2016). “Neoclassica - A Multilingual Domain Ontology”. In: *Computational History and Data-Driven Humanities: Second IFIP WG 12.7 International Workshop, CHDDH 2016, Dublin, Ireland, May 25, 2016, Revised Selected Papers*. Ed. by Bojan Bozic et al. Cham: Springer International Publishing, pp. 41–53. ISBN: 978-3-319-46224-0. DOI: 10.1007/978-3-319-46224-0_5. URL: https://doi.org/10.1007/978-3-319-46224-0_5.
- Johnson, Justin M. and Taghi M. Khoshgoftaar (2019). “Survey on Deep Learning with Class Imbalance”. In: *Journal of Big Data* 6.1, p. 27. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0192-5. URL: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5> (visited on 12/16/2019).
- Lang, Sabine and Björn Ommer (2018). “Attesting Similarity: Supporting the Organization and Study of Art Image Collections with Computer Vision”. In: *Digital Scholarship in the Humanities* 33.4, pp. 845–856. ISSN: 2055-7671, 2055-768X. DOI: 10.1093/linc/fqy006. URL: <https://academic.oup.com/dsh/article/33/4/845/4964861> (visited on 12/16/2019).
- Lenci, Alessandro (2018). “Distributional Models of Word Meaning”. In: *Annual Review of Linguistics* 4.1, pp. 151–171. ISSN: 2333-9683, 2333-9691. DOI: 10.1146/

`annurev-linguistics-030514-125254`. URL: <http://www.annualreviews.org/doi/10.1146/annurev-linguistics-030514-125254> (visited on 12/16/2019).

Mikolov, Tomas et al. (2013). “Efficient Estimation of Word Representations in Vector Space”. In: arXiv: 1301.3781 [cs]. URL: <http://arxiv.org/abs/1301.3781> (visited on 12/16/2019).

Sales, Juliano Efson et al. (2018). “Indra: A Word Embedding and Semantic Relatedness Server”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair) et al. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

MULTIMODALER BEDEUTUNGSTRANSFER VOM TEXT ZUM BILD. GRANULARE BILDKLASSIFIKATION DURCH VERTEILUNGSSEMANTIK.

Simon Donig 

Lehrstuhl für Digital Humanities
Universität Passau, Deutschland
simon.donig@uni-passau.de

Maria Christoforaki

Lehrstuhl für Data Science
Institut für Informatik
Universität St.Gallen, Schweiz
maria.christoforaki@unisg.ch

Bernhard Bermeitinger 

Lehrstuhl für Data Science
Institut für Informatik
Universität St.Gallen, Schweiz
bernhard.bermeitinger@unisg.ch

Siegfried Handschuh

Lehrstuhl für Data Science
Institut für Informatik
Universität St.Gallen, Schweiz
siegfried.handschuh@unisg.ch

Dezember 2019

1 Einleitend

In den letzten Jahren hat die Verwendung von Bildklassifizierungsverfahren wie neuronalen Netzwerken auch im Bereich der historischen Bildwissenschaften und der *Heritage Informatics* weite Verbreitung gefunden (Lang und 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, Bermeitinger u. a. 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* (CNN) genutzt, dessen Ausgabe im Trainingsprozess, anders als herkömmlich, nicht auf einer Serie flacher Textlabel be-

ruht, 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).

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 mithilfe des DSMs in natürlichsprachige Wörter zurückgewandelt wird. Da wir nach reichereren Repräsentationen im Zuge dieses Vorgangs suchen, wählen wir die fünf nächsten Nachbarn aus (*Top-5 Nearest Neighbors*). Als Ähnlichkeitsmaß legen wir die Kosinusähnlichkeit zwischen dem vorhergesagtem Vektor und jenem Vektor zugrunde, der dem ursprünglich dem Bild von uns zugewiesenen Textlabel entspricht. Wir gehen davon aus, dass ein Bild korrekt klassifiziert wurde, wenn das Goldlabel unter diesen 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.

2 Aufbau des Experiments

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

2.1 Textkorpus

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-NLP-Verfahren angewandt, darunter Tokenisierung, Satz- und Worttrennung, die Normalisierung von Zahlenwerten und die Erkennung von benannten Entitäten (*Named Entity Recognition, NER*). Da wir retrodigitalisiertes Material aus verschiedenen Quellen nutzen, implementieren wir manuelle Korrekturen für die häufigsten der vorkommenden Fehler (etwa Ligaturen wie II, die als U durch das OCR-Verfahren fehlinterpretiert werden). Eine weitere Ebene der Vorverarbeitung besteht aus inhaltsbezogenen Augmentierungen. Insbesondere normalisieren wir zusammen gesetzte Wörter und Synonyme gemäß einer spezifizierten Liste, die anhand einer Ontologie, der *Neoclassica-Ontologie* (Donig, Christoforaki und Handschuh 2016) zusammengestellt sind. Dies resultiert in einem Korpus von insgesamt 3 067 237 Wörtern

aus 107 518 Wortgrundformen. Das DSM wird von uns mithilfe des *Indra*-Frameworks (Sales u. a. 2018) sowie *Gensim* (Řehůřek und Sojka 2010) erzeugt¹.

2.2 Bildkorpus

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.

2.3 Kombiniertes Korpus

Da es sich um ein *Proof-of-Concept*-Experiment handelt, kommt zum Zweck des *Rapid Prototypings* ein an die VGG-Architektur (Krizhevsky, Sutskever und Hinton 2012) angelehntes, „simples“ neuronales Netzwerk zum Einsatz³. Die Unabhängigkeit der Trainings- und Evaluationsbeispiele wird durch eine 5-fache Kreuzvalidierung auf 80 % des Korpus garantiert, wovon wiederum 80 % für das Training und 20 % für das Testen benutzt werden. Die übrigen 20 % werden als Evaluationsmenge behandelt. Die reportierten Ergebnisse basieren auf diesen, dem Klassifikator unbekannten 20 %.

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

¹Die Erzeugung erfolgt mit einer Vektorgroße von 50, einer Worfenergröße von 10 und einer minimalen Wortzahl von fünf erzeugt. Als *Word2Vec*-Modell kommt *Skipgram* (Mikolov u. a. 2013) mit *Negative Sampling* zum Einsatz.

²Das Korpus wurde aus den Sammlungen des Metropolitan Museum (New York), des Victoria & Albert Museum (London), der Wallace Collection (London) sowie mehreren zeitgenössischen Musterbüchern zusammengestellt.

³Das Netzwerk besteht aus drei Convolutional-Blöcken mit jeweils zwei Convolutional-Layer mit 32, 64 und 64 Filter der Größe 3×3 . Nach jedem Block folgt ein Maximum-Pooling-Layer der Größe 2×2 , 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 Frameworks *Keras* (<https://keras.io>) und *TensorFlow* (<https://tensorflow.org>) implementiert. Beim Training wird der durchschnittliche absolute Fehler durch die Optimierungsfunktion *RMSprop* minimiert.

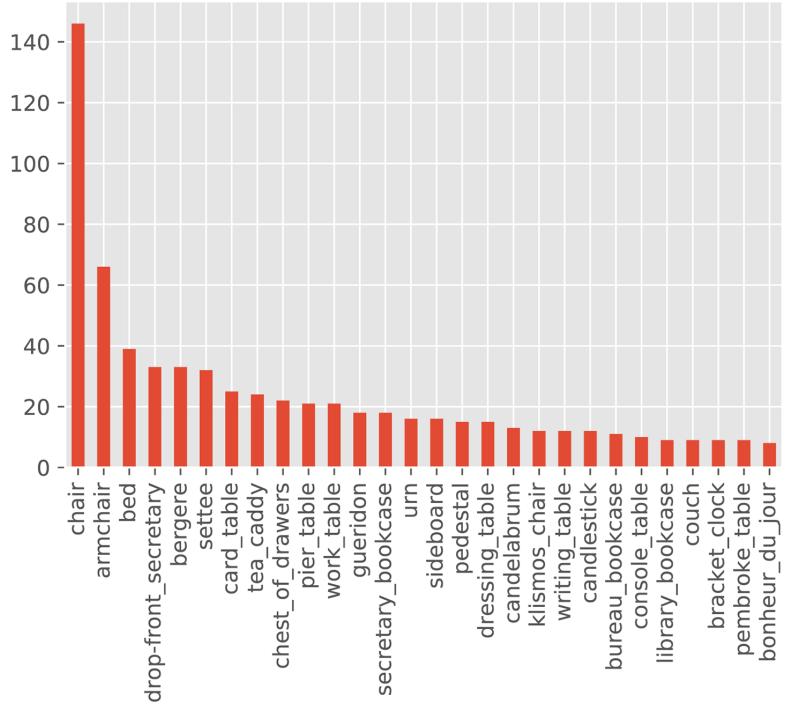


Abbildung 1: Verteilung der Klassen im Bildkorpus.

3 Ergebnisse

Die Top-5-Richtig-Positiv-Rate (*true-positive*) beträgt 0,59. Das bedeutet, dass das Goldlabel in 59 % der Fälle unter den fünf nächsten Nachbarn erscheint.

Das mathematische Qualitätskriterium gibt für sich genommen jedoch nur einen Teil des Gesamtbilds wieder. Wir führen deshalb zugleich eine qualitative Analyse der Ergebnisse in der Evaluationsteilmenge durch.

Eine Reihe von richtig-positiven Ergebnissen zeigen, 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 (Ankleide-tisch, 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 dezidiert 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



Abbildung 2: Abweichungen bei der Klassifizierung desselben Objekts (Victoria & Albert Museum 2019b).

konsistent ist (die Top-4 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 ausgelöst 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.

Während die Label in den bisher betrachteten Fällen die taxonomischen Beziehungen reflektieren und alle den aus der Ontologie abgeleiteten Zielwörtern entstammen, zeigt Abb. 4, dass das Verfahren auch aus sich selbst, rein datenzentriert Label generieren kann. Die abgebildete Kratervase wurde manuell als Urne (*urn*) annotiert. 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“ oder „bell-crater“ (mit und ohne Bindestrich) existieren, um diese Art von Artefakten zu beschreiben. „Ovoid“ bezieht sich demgegenüber wohl auf die Eierstabdekorations des oberen Wulstes, die oft mit diesem Adjektiv beschrieben wird. Diese Ornamentik scheint zugleich die Assoziation zur Rosette (*Patera*) mitbedingt zu haben. Auf diese Weise erscheint das Zielwort *patera_element* unter den Top-5, obwohl im Bildkorporus ausschließlich ganze Artefakte, nicht aber deren Dekor annotiert wurden.

Nicht auszuschließen ist hier zudem ein Effekt des visuellen Klassifikators, wie auch Abb. 5 zeigt. Die Fehlklassifikation des Objekts, eines Nähtischchens, führte zu konstistenten Zuschreibungen im Bereich der Sitz- und Liegemöbel. Betrachtet man die



Abbildung 3: Abweichungen bei der Klassifizierung desselben Objekts (Victoria & Albert Museum 2019b).

äußere Form des Artefakts auf einer abstrakteren Ebene, kann man eine visuelle Nähe zu z. B. einem (*Double-*) *Camel-back Sofa* durchaus nachvollziehen.

4 Vergleich zu einem herkömmlichen CNN

Um die Unterschiede zwischen beiden Zugängen besser abschätzen zu können, führen wir weiter in Vergleichsexperiment durch, bei dem dasselbe CNN⁴ wie im vektorbasierten Verfahren für eine herkömmliche Klassifizierung mit flachen Labeln herangezogen wird.

Um die Unterschiede der beiden Zugänge (herkömmliche Klassifizierung mit flachen Labeln/vektorbasiertes Verfahren) besser abschätzen zu können, stellen wir im Folgenden einen Vergleich zentraler Metriken für beide Zugänge vor.

Wie in Tabelle 1 ersichtlich⁵, ist nicht nur die Top-1-Treffergenauigkeit im Fall der Klassifizierung mit Vektoren besser, sondern es sind auch die übrigen Metriken vergleichbar gut. Durch den hier vorgeschlagenen Zugang wird also nicht nur die Treffergenauigkeit verbessert, sondern er liefert zugleich eine reichhaltigere Beschreibung des Bildes.

⁴Das verwendete CNN ist identisch mit dem im Vektor-Experiment verwendeten Netzwerk, mit Ausnahme des letzten Layers, der aufgrund der Labelklassifikation nun 28 statt 50 Ausgaben besitzt und die Aktivierungsfunktion *Softmax* statt einer linearen.

⁵Die englischen Begriffe für bessere Verständlichkeit: Treffergenauigkeit (*Accuracy*), Genauigkeit (*Precision*), Trefferquote (*Recall*), Sensitivität (*True-Positive Rate*)



Abbildung 4: Eine Sèvres-Kopie der Medici-Vase stößt die Klassifizierung mit assoziativen Labeln an. (Victoria & Albert Museum 2019a)

Metrik	Vektor	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-Maß	0,26	0,25
Top-5-Sensitivität	0,59	0,73
Top-5-Falsch-Positiv-Rate	0,41	0,27

Tabelle 1: Vergleich zentraler Metriken beider Zugänge.

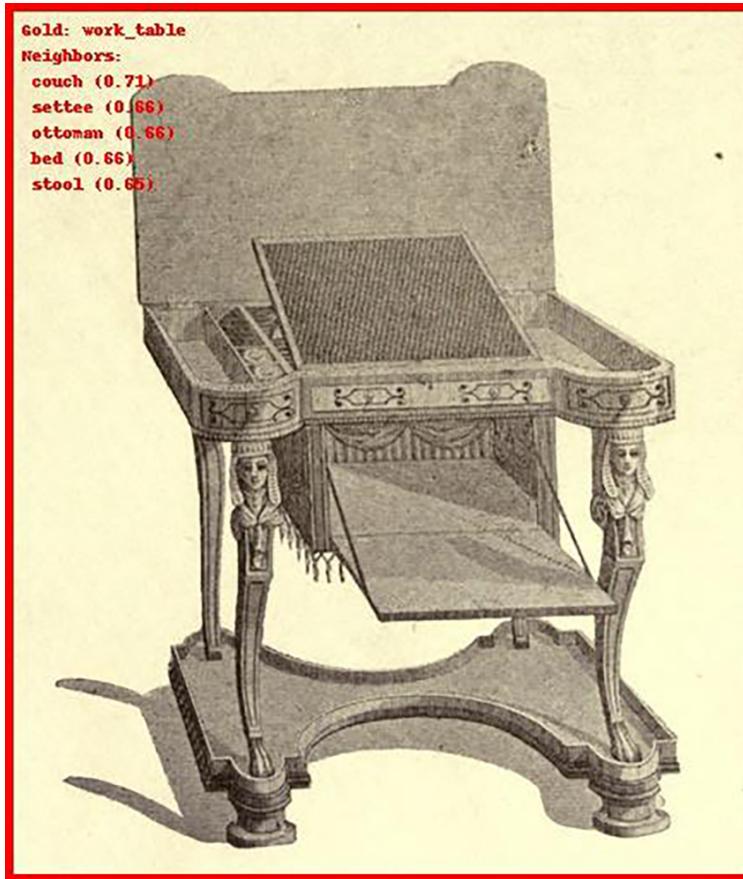


Abbildung 5: Fehlklassifikation eines Nähtischs in ein Wortumfeld aus der Sitzmöbel-Hierarchie. (Sheraton, Hayden und Bell 1910, S. 88)

5 Schlussfolgerungen und Ausblick

In dem hier vorgeschlagenen Beitrag stellen wir ein neues, multimodales Verfahren für die Klassifikation von Bildinhalten vor, das auf der Kombination von NLP-Methoden mit Bildklassifizierungsverfahren beruht. Ziel ist, Objekte nicht alleine nach einem Schema flacher Label, sondern in einer kontextgerechteren Weise zu klassifizieren, wobei dieser Kontext von einschlägigen historischen Domänepublikationen 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 erreichen wir eine Treffergenauigkeit von 0,59. Als nächsten Schritt möchten wir mit einem komplexeren CNN und einem ausgeweiteten Bildkorpus trainieren (um bekannte Probleme wie *Overfitting* zu reduzieren). Das Ver-

gleichsexperiment mit einem herkömmlichen, auf flachen Labeln beruhenden Zugang zeigt, 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 Zielwörter 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 Zielwörter zu erstellen. In ähnlicher Weise ziehen wir in Betracht, benannte Entitäten zu URIs zusammenzufassen. Das würde uns erlauben spezifische Entitäten (z. B. Werkstätten, Ebenisten und 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.

Literatur

- Donig, Simon, Bernhard Bermeitinger u. a. (2019). „Vom Bild Zum Text Und Wieder Zurück“. In: *Digital Humanities: Multimedial & Multimodal - Konferenzabstracts*. DHd2019. Frankfurt am Main, Germany. URL: <https://www.researchgate.net/publication/332275547>.
- Donig, Simon, Maria Christoforaki und Siegfried Handschuh (2016). „Neoclassica - A Multilingual Domain Ontology“. In: *Computational History and Data-Driven Humanities: Second IFIP WG 12.7 International Workshop, CHDDH 2016, Dublin, Ireland, May 25, 2016, Revised Selected Papers*. Hrsg. von Bojan Bozic u. a. Cham: Springer International Publishing, S. 41–53. ISBN: 978-3-319-46224-0. DOI: 10.1007/978-3-319-46224-0_5. URL: https://doi.org/10.1007/978-3-319-46224-0_5.
- Johnson, Justin M. und Taghi M. Khoshgoftaar (2019). „Survey on Deep Learning with Class Imbalance“. In: *Journal of Big Data* 6.1, S. 27. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0192-5. URL: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5> (besucht am 16.12.2019).
- Krizhevsky, Alex, Ilya Sutskever und Geoffrey E. Hinton (2012). „ImageNet Classification with Deep Convolutional Neural Networks“. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. (Lake Tahoe, Nevada). NIPS’12. USA: Curran Associates Inc., S. 1097–1105. URL: <http://dl.acm.org/citation.cfm?id=2999134.2999257> (besucht am 16.12.2019).
- Lang, Sabine und Björn Ommer (2018). „Attesting Similarity: Supporting the Organization and Study of Art Image Collections with Computer Vision“. In: *Digital Scholarship in the Humanities* 33.4, S. 845–856. ISSN: 2055-7671, 2055-768X. DOI: 10.1093/linc/fqy006. URL: <https://academic.oup.com/dsh/article/33/4/845/4964861> (besucht am 16.12.2019).

- Lenci, Alessandro (2018). „Distributional Models of Word Meaning“. In: *Annual Review of Linguistics* 4.1, S. 151–171. ISSN: 2333-9683, 2333-9691. DOI: 10.1146/annurev-linguistics-030514-125254. URL: <http://www.annualreviews.org/doi/10.1146/annurev-linguistics-030514-125254> (besucht am 16.12.2019).
- Mikolov, Tomas u. a. (2013). „Efficient Estimation of Word Representations in Vector Space“. In: arXiv: 1301.3781 [cs]. URL: <http://arxiv.org/abs/1301.3781> (besucht am 16.12.2019).
- Řehůřek, Radim und Petr Sojka (2010). „Software Framework for Topic Modelling with Large Corpora“. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, S. 45–50.
- Sales, Juliano Efson u. a. (2018). „Indra: A Word Embedding and Semantic Relatedness Server“. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Hrsg. von Nicoletta Calzolari (Conference chair) u. a. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.
- Sheraton, Thomas, Arthur Hayden und J. Munro Bell (1910). *The Furniture Designs of Thomas Sheraton*. London: Gibbings and Co., Ltd. URL: <https://archive.org/details/furnituredesigns00sheruoft>.
- Victoria & Albert Museum (2019a). *Vase [Sèvres Copy of the Medici Vase], Paris, 1813. Ascension Number 396-1874*. URL: <http://collections.vam.ac.uk/item/08978/vase-sevres-porcelain-factory/>.
- (2019b). *Writing Table, Neuwied, Workshop of David Roentgen ca. 1774-1780. Ascension Number: 1059:1 to 9-1882*. URL: <http://collections.vam.ac.uk/item/0117298/writing-table-roentgen-david/>.