
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 classification 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.

The conducted experiment is based on the collation of two corpora: one textual 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 assume 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 better than the label-based version in some aspects.

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.

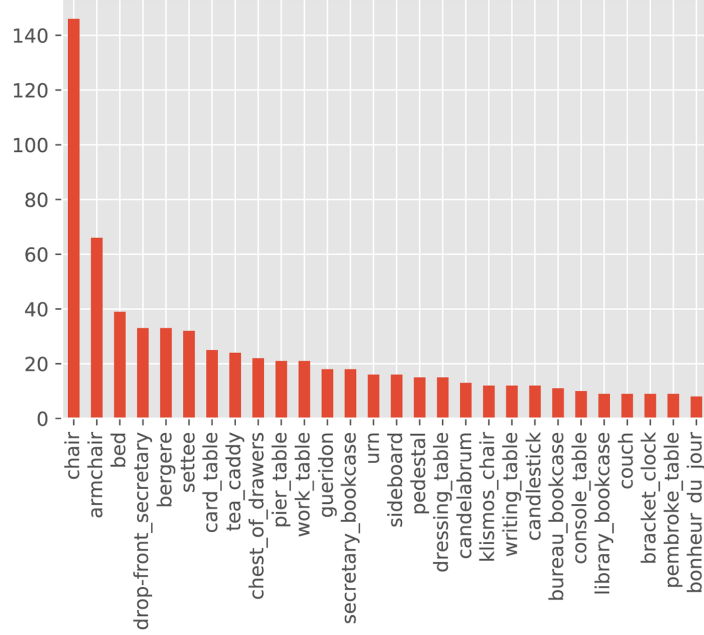


Figure 1: Class distribution of the image corpus

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 0.59, meaning that the gold label from the annotations is in 59 % 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 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 between the different approaches 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-

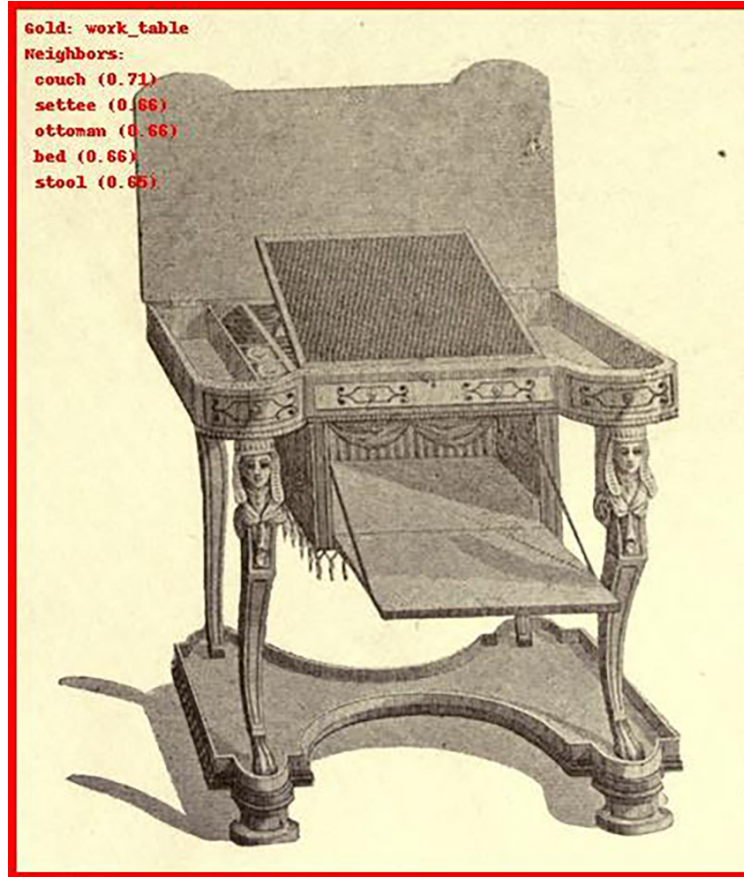


Figure 5: Misclassification of a sewing table in the proximity of seating furniture hierarchy.

Metric	Vector-based	Label-based
Top-1 Accuracy	0.50	0.40
Top-1 Precision	0.32	0.29
Top-1 Recall	0.25	0.26
Top-1 F1-Score	0.26	0.25
Top-5 True-Positive-Rate	0.59	0.73
Top-5 False-Positive-Rate	0.41	0.27

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 0.59.

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. (Mar. 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 (Dec. 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 (Dec. 1, 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/llc/fqy006. URL: <https://academic.oup.com/dsh/article/33/4/845/4964861> (visited on 12/16/2019).
- Lenci, Alessandro (Jan. 14, 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. (Sept. 6, 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. (May 7–12, 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.