



UvA-DARE (Digital Academic Repository)

A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections

Smits, T.; Wevers, M.

DOI

[10.1093/llc/fqad008](https://doi.org/10.1093/llc/fqad008)

Publication date

2023

Document Version

Final published version

Published in

Digital Scholarship in the Humanities

License

CC BY-NC

[Link to publication](#)

Citation for published version (APA):

Smits, T., & Wevers, M. (2023). A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digital Scholarship in the Humanities*, 38(3), 1267-1280. Advance online publication. <https://doi.org/10.1093/llc/fqad008>

General rights



It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)

A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections

Thomas Smits ^{1*}, Melvin Wevers ²

¹Faculty of Arts, University of Antwerp, Antwerp, Belgium

²Department of History, Faculty of Humanities, University of Amsterdam, Amsterdam, The Netherlands

*Correspondence: Thomas Smits, Grote Kauwenberg 18, 2000 Antwerpen, Belgium. E-mail: thomas.smits@uantwerpen.be

Abstract

Until recently, most research in the Digital Humanities (DH) was monomodal, meaning that the object of analysis was either textual or visual. Seeking to integrate multimodality theory into the DH, this article demonstrates that recently developed multimodal deep learning models, such as Contrastive Language Image Pre-training (CLIP), offer new possibilities to explore and analyze image–text combinations at scale. These models, which are trained on image and text pairs, can be applied to a wide range of text-to-image, image-to-image, and image-to-text prediction tasks. Moreover, multimodal models show high accuracy in zero-shot classification, i.e. predicting unseen categories across heterogeneous datasets. Based on three exploratory case studies, we argue that this zero-shot capability opens up the way for a multimodal turn in DH research. Moreover, multimodal models allow scholars to move past the artificial separation of text and images that was dominant in the field and analyze multimodal meaning at scale. However, we also need to be aware of the specific (historical) bias of multimodal deep learning that stems from biases in the training data used to train these models.

1 Introduction

Until the mid-2010s, Digital Humanities (DH) research had a clear focus on text analysis (Champion, 2017; Manovich, 2020). An important driver behind this focus on text was the large-scale digitization of historical sources, for example, by Google Books, which transferred out-of-copyright books to a digital format (Leetaru, 2008). In the early 2000s, Optical Character Recognition (OCR), a second technological innovation, sped up the transformation of digitized materials into machine-readable text. In the same period, rapid advancements in Natural Language Processing (NLP) enabled new ways of studying digital and turned-digital textual sources. Humanities scholars followed suit and started applying NLP methods to study cultural and historical phenomena in books, newspapers, and other digitized sources (Bingham, 2010). While some described the combination of digitization and computational methods as a ‘practical revolution’ (Nicholson, 2013), others argued that it fundamentally

altered the methodological foundation of the humanities. Developed in the early 2000s, Moretti’s (2000, 2015) concept of ‘distant reading’ has become the primary methodological framework to describe the ‘humanities 3.0’ (Bod, 2013).

Whether we have witnessed such a practical or methodological revolution is not the focus of this article. We do want to emphasize that by focusing on text, DH scholars long overlooked a defining aspect of modernity: the increasing importance of visual forms of representation (Paul, 2016). Starting with illustrations in the nineteenth century and culminating in the almost incessant streams of photographs that are uploaded to social media, modern media have increasingly included visual information. In the last 10 years, the rapid development of computer vision (CV) paved the way for the computational study of the millions of images in digitized and born-digital collections. We noted that convolutional neural networks (CNNs), a specific type of deep learning model, can be used to identify trends in

large collections of images (Wevers and Smits, 2020). Building on Moretti's concept, Arnold and Tilton (2019) argued that the combination of the digital availability of images and new computational methods should lead to a new 'distant viewing' methodology.

Despite the usefulness of NLP and CV methods for humanities, there are conceptual issues with approaching the textual and visual as separate entities. Work in semiotics, visual culture studies, and multimodality theory demonstrates that the meaning of texts and images in modern media, such as the newspaper, magazine, or the Internet, cannot be understood separately, i.e. as only visual or only textual. In 1961, French semiotician Barthes (1961) already noted that 'the photograph is not an isolated structure; it is in communication with at least one other structure, namely the text'. In 1977, Sontag argued that photographs needed texts to anchor them in the plurality of meaning that they could convey. In visual culture studies, Mitchell (2005) provocatively asserted that 'there are no visual media'. All images are 'riddled ... with language'. More recently, multimodality theory shed light on the interaction between text and images in different 'semiotic modes' (Hiippala, 2021). The concept of meaning multiplication describes the idea that the meaning of an image-text combination—a photograph and a caption, for example—is not the same if we would see the text and image independently from each other (Bateman, 2014).

In DH, the separation of texts and images was never based on theoretical considerations. Rather, it was the direct result of the two practical revolutions described above. After all, NLP and CV techniques can only be used to analyze one mode at a time. In fact, some scholars warned that the possibilities of NLP would result in a neglect of visual content, such as the photographs, cartoons, sketches, and maps of digitized newspapers (Wijffes, 2017). Starting from the analysis of images, van Noord (2022) noted that CV techniques are still unable to bridge the 'semantic gap' (Smeulders *et al.*, 2000) between what computational methods can extract from visual data and what these data mean to a user. In other words, we need text if we want to understand what images mean.

Building on recent work that seeks to integrate multimodality theory into DH research (Hiippala, 2021; Smits and Ros, 2021), this article argues that multimodal deep learning models can be used to analyze image-text combinations at scale. Trained to learn which images and texts belong together, they can be applied to a wide variety of text-to-image, image-to-image, and image-to-text prediction tasks (Jia *et al.*, 2021; Radford *et al.*, 2021). In contrast to monomodal CV models, which often must be trained on additional data relevant to the task at hand, multimodal models show a high performance 'in the wild' on various tasks

applied to heterogeneous datasets for which they were not optimized during training. Put differently, these models produce reliable results even when applied to data that they did not encounter during training. We demonstrate that this 'zero-shot' capability makes them especially useful for analyzing the representation of complex multimodal concepts in small and large visual archives.

Following sections on multimodal models and their potential for DH research, this article describes three research projects where we applied the multimodal model Contrastive Language Image Pre-training (CLIP) to explore, enrich, and analyze three different visual cultural heritage collections. In the first project, we used the model to identify images of exterior/interior scenes in a set of 42,000 late-nineteenth-century digitized magic lantern slides. The second project employs CLIP to study the representation of the family in a collection of around 38,000 illustrations in Dutch children's books published between 1840 and 1940. Finally, we describe how CLIP can be used to quickly label a large collection of historical press photographs. After being checked by crowd-workers, this labeled data is used to fine-tune the CLIP model and as input for transfer learning of a scene detection model.

Based on the case studies, we argue that multimodal models have the potential to cause a multimodal turn in DH research. Instead of spending time and resources on labeling data, training, or fine-tuning, these models allow scholars to easily analyze multimodal meaning on domain-specific data. They also enable researchers to move past the artificial separation of text and images that characterized DH research thus far. A word of caution, however, is at place. Researchers have to scrutinize the output of multimodal models when applied to (historical) cultural data. The final section of this article shows how biases in the data used to train these models might impact the output generated by multimodal models.

2 Contrastive multimodal models

In the last 10 years, the construction of large-scale annotated datasets, such as ImageNet (Russakovsky *et al.*, 2015), Google Open Images (Kuznetsova *et al.*, 2020), and MS COCO (Lin *et al.*, 2014) fueled the rapid development of CV. Because of the required human annotation, producing these annotated datasets was time-consuming and expensive. It took around 50,000 different workers 2 years to populate ImageNet (Reese and Heath, 2016). In a similar project, workers spent around 70,000 hours annotating 2.5 million instances of the ninety-one classes of MS COCO (Lin *et al.*, 2014). Following the machine learning adage 'there is no data like more data', CV experts claimed

that the time and money needed to produce labeled data was the primary bottleneck for developing better-performing models. As a result, they started to look for ways to use unlabeled data to train CV models.

In early 2021, building on the success of GPT-3 and other pre-training methods in NLP, the Microsoft-backed research laboratory OpenAI developed a multimodal machine learning model that could learn visual concepts from unlabeled visual data. While traditional CV models are trained to identify a limited set of objects and/or persons on images, OpenAI’s CLIP is optimized to connect images and texts using 400 million image–text combinations as training data (Radford *et al.*, 2021). CLIP and similar models, such as Google’s Align (Jia *et al.*, 2021), rely on two encoders to turn both images and corresponding texts into vectors. During training, the models are fitted to maximize the cosine similarity of the vectors of the original image–text pairs and minimize the cosine similarity between all the other possible combinations. This learning objective is a contrastive—the C in CLIP—loss task (Fig. 1). In contrast to other multimodal models, such as VirTex (Desai and Johnson, 2021), contrastive multimodal models are not trained to predict the captions of images but only to estimate if an image–text pair likely occurred in the original dataset (Jia *et al.*, 2021; Radford *et al.*, 2021).

The developers of CLIP argue that their model learns to connect visual and language representations of the image and text pairs (Radford *et al.*, 2021). In other words, multimodal models learn which visual elements are good predictors of specific textual elements and vice versa. As a result, they can be used to turn any set of images into embeddings and calculate the cosine similarity between them and the embedding of any textual prompt, which the user provides to the model

(Fig. 2). The authors of CLIP demonstrate this ‘zero-shot’ capability by testing its performance on twenty-seven different CV benchmarks. Without task-specific training, the model matches, or (slightly) improves the performance of a purposefully trained ResNet-50 CV model on sixteen of them, including well-known ones, such as ImageNet, Caltech 101, and PASCAL VOC.

3 How can DH researchers use multimodal models?

This section starts by describing how multimodal models can be used for three types of retrieval tasks (text-to-image, image-to-text, and image-to-image) to query, enrich, and analyze visual digital heritage collections. At the end of the section, we describe how the application of these models can help to answer humanities research questions.

For all three tasks, we start by applying the trained model to extract the embeddings of a collection of images. For the text-to-image task, we connect textual prompts, which are turned into embeddings by the model, to the embeddings of the collection’s images. Subsequently, we can look for images using generic prompts such as ‘animals’, but also more specific ones such as ‘lion’ or ‘dog’ (Fig. 3). We can retrieve images of specific scenes, a church service, a wedding, demonstrations, or a soccer match; places, such as landmarks or cities; types of persons, soldiers, priests, and teachers; or even specific persons, like Queen Victoria or former president Barack Obama. However, we can also identify multimodal concepts that are hard to describe in purely visual terms. For example, we can retrieve images of abstract concepts, such as love or anger (Fig. 4). For this text-to-image retrieval task, it is important to note that multimodal is particularly sensitive

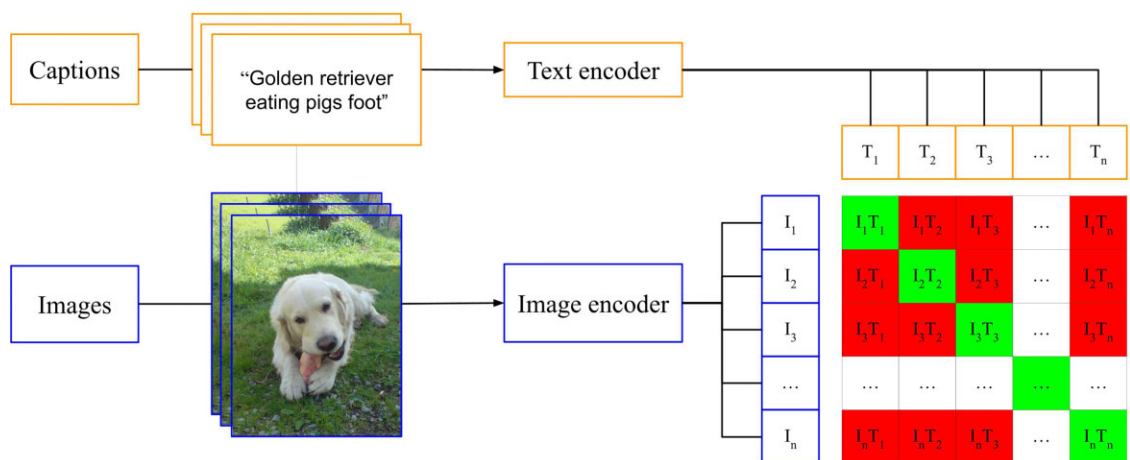


Figure 1. A visual representation of contrastive pre-training. The figure is based on an image that can be found in a blog about CLIP by OpenAI (<https://openai.com/blog/clip/>)

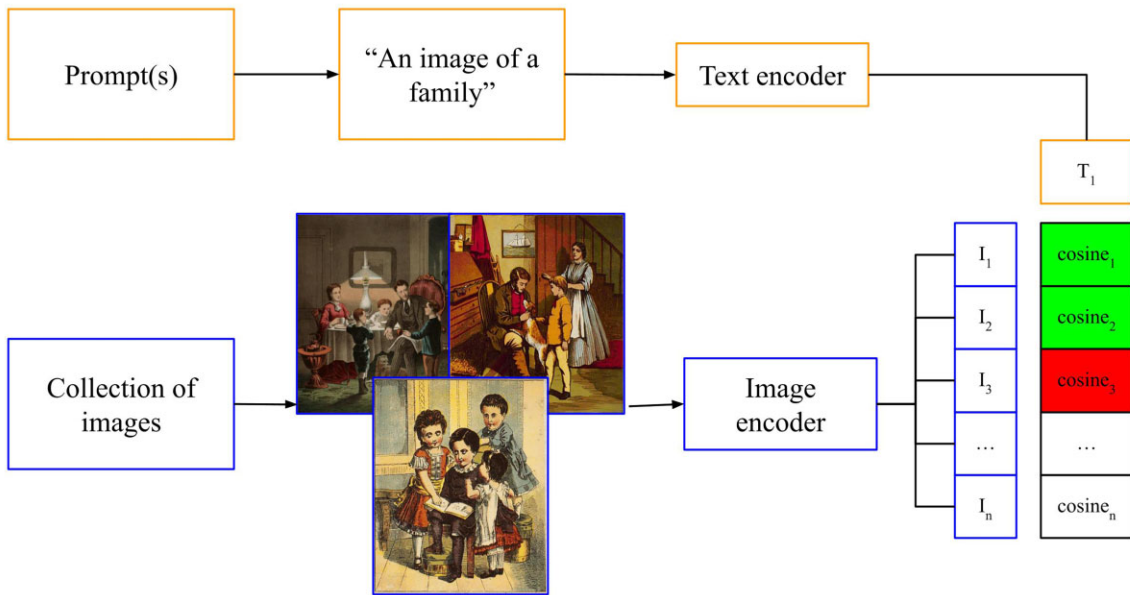


Figure 2. A visual representation of zero-shot prediction on a historical collection of images using prompts. The figure is based on an image that can be found in a blog about CLIP by OpenAI (<https://openai.com/blog/clip/>)

to text on images (Radford *et al.*, 2021). For example, the word ‘collie’ above the image of the second dog in Fig. 3 will lead the model to connect this image more easily to the prompt ‘dog’. The same would go for the prompt ‘bakery’ and a photograph where the word ‘bakery’ is visible on the storefront.

Researchers should carefully consider the prompts they use to query the model. CLIP’s developers underline this by noting how the phenomena of polysemy—the capacity of a word to have multiple-(related) meanings—might be a hurdle in the construction of prompts (Radford *et al.*, 2021). For example, the CV dataset ImageNet contains two classes of ‘crane’: one referring to the bird and the other to the machine. This fact is inconsequential to a traditional CV model, which, in this case, connects two classes to different images. However, a contrastive multimodal model will turn the word crane into a single embedding and, as a result, will connect images of both birds and machines to it. Next to polysemy, we note that other semantic phenomena are also relevant here. For example, homonymy, where the similarity between words cannot be explained etymologically, also leads to problems in the construction of prompts. In fact, ‘crane’ is a homonym rather than a polyseme. A ‘boxer’, the second example that the developers of CLIP mention in their paper, is a polyseme: the German dog breed boxer was named after the English word boxer. Figures of speech might lead to similar problems in the construction of prompts. For example, metonymy (‘a pair of hands’/ ‘twenty sails’) and metaphors (‘swallowing a bitter

pill’) might lead the model to suggest literal visual representations which lack the metonymical layer. The developers of CLIP note that this problem can be (partly) solved by adding context to the prompts. For example, ‘crane, a type of bird’ instead of ‘crane’. (Radford *et al.*, 2021).

Multimodal models are not only able to connect texts to images but can also be used for other ‘cross-modality matching [and] retrieval tasks’ (Jia *et al.*, 2021). While they cannot generate captions, they can calculate the cosine similarity between an image and a user-defined set of textual inputs. Researchers could use this ‘image-to-text’ task to quickly and accurately generate meta-data for a visual collection. Connecting the prompts ‘an oil painting’, ‘an illustration’, and ‘a photograph’ to images, we could chart the distribution of visual media in a collection.

Just like the latent space of textual models display geometric regularities—‘queen is to king what woman is to man’ in the famous example—multimodal models are also known to represent such regularities between image and text embeddings during training (Jia *et al.*, 2021). In practice, we can add or subtract a text query to or from an image and then use the resulting embedding to retrieve relevant images via a cosine similarity score. For example, a query for the Eiffel Tower plus a query with the word ‘snow’ will turn up images of the famous Parisian landmark in the winter (Jia *et al.*, 2021).

Because multimodal models are trained to connect text to images, similar images are located close to each



Figure 3. Top-4 cosine similarity scores (0.273/0.254/0.253/0.251) for the prompt ‘a dog’ and a random sample of 10.000 of the 42.000 magic-lantern slides. Reproduced by permission via Lucerna Magic Lantern Web Resource. Left to right: ‘Hond’ (item 5106528), slide 12 of ‘Ons huis’ (c. 1920). Digital image uploaded by Sarah Dellmann (2016); ‘Collie’ (item 5040076), slide 24 of ‘Instantaneous studies: animals’ (c. 1891). Image uploaded by the Manchester Museum (2019); ‘Tearem Pincher, Esq.’ (item 5069469) and ‘Sir Robert Pincher’ (item 5069468), slides 3 and 4 of ‘Comical Cats and Dogs’ (1900–1907). Image uploaded by Lucerna (2019)

other in the embedding space. This proximity means that we cannot only search for images with textual prompts but also with images: an image-to-image retrieval task. Provided with an image of London Bridge, a multimodal model might return more visual representations of the same landmark, other bridges, and London (if they are in the dataset). We can also use multimodal models for several similarity and clustering tasks that DH researchers commonly use. Multimodal models can retrieve (near) duplicates in a visual collection in a manner similar to Imagededup, a popular library for this specific task (Jain *et al.*, 2019/2020). In addition to (near) duplicates, they can be used to identify visually similar images to an input image (Seguin *et al.*, 2017; Wevers and Lonij, 2017). Similar to Yale DH Lab’s popular PixPlot method, embeddings generated by contrastive models can be used to cluster images into groups of similar images (Duhaime and Leonard, 2017/2020). Finally, as we show in the third case study, we can train a classifier on top of the embeddings. Like the Tensorflow for Poets method,

which trains a classifier on top of ImageNet features using a small number of training images, we can thus use the embeddings generated by a multimodal model to produce a visual classification model quickly.¹

Using contrastive multimodal models trained on millions of image–text pairs to find duplicates might seem like using a sledgehammer to kill a fly. As a rule, researchers should turn to simple solutions to solve simple tasks. Concerning the image similarity task, hashing algorithms might be able to do the same job as a multimodal model for a fraction of the compute power. However, it is important to note that the embeddings of multimodal models can be used for various tasks. Researchers could re-use the same embeddings to identify duplicates; cluster images; add meta-data to a collection; identify objects, persons, or scenes; and to train specific classifiers. Moreover, the same embeddings can be used by different groups: a museum could use them to add meta-data to a visual collection, make them available to scholars for research, and to ordinary users interested in retrieving



Figure 4. Top-4 cosine similarity scores (0.257/0.249/0.249/0.246) for the prompt ‘a couple in love’ and a random sample of 10.000 of the 42.000 magic-lantern slides. Reproduced by permission via Lucerna Magic Lantern Web Resource. Left to right: ‘Who knows how I long to kiss you?’ (item 5009448), slide 15 of ‘Who knows? Who cares?’ (1910). Image uploaded by Lucerna (2008); ‘Oh, sad was my heart when we sobbed our good-bye’ (item 5005994), slide 3 of ‘Angus MacDonald: new series’ (1912). Image uploaded by Ludwig Vogl-Bienek (2006); ‘And out on the shingle, we leap in our glee’ (item 5011569), slide 4 of ‘Mona: new series’ (1905). Image uploaded by Lucerna (2008); ‘Tis Angus my own’ (item 5006007), slide 16 of ‘Angus MacDonald: new series’ (1912). Image uploaded by Ludwig Vogl-Bienek (2006)

specific images from a collection. This wide variety of tasks, uses, and users might justify the application of multimodal models in many cases, despite the fact that it might seem like a rather generic approach.

Multimodal models provide a new kind of bottom-up access to large visual collections. Without having to (manually) add metadata, they allow researchers to search for a wide range of visual concepts in large collections of images in the same way that OCR allowed keyword search in large collections of texts. What kind of questions can be answered with this new possibility? While this question depends on the interest of the researcher, we note that humanities scholars have always been interested in the visual representation of complex phenomena and concepts over time. Think, for example, of war, gender, love, or piety. Multimodal models make it easier than ever to quickly identify visual representations of these concepts in a large collection and analyze how their meaning changes over time. Here,

multimodal models can not only be used to study *what* is visually represented but also *how* this is done. Visual styles and other aesthetical elements can be easily added to search queries. The three case studies below provide examples of how this kind of research might look.

4 Case studies

In the three case studies outlined in this article, we apply the multimodal model CLIP to query, enrich, and analyze three different visual heritage collections and answer questions of interest to humanities scholars. The cases showcase the possibilities of contrastive multimodal models for DH research and highlight methodological challenges. The first two cases use the same methodological setup. After applying CLIP to turn a set of images into embeddings, we provide it with different prompts to identify visual concepts in an

extensive visual collection. In the third case study, we rely on CLIP to generate labels for unlabeled data to quickly create new training data for a CV model.

4.1 Outdoor/indoor in nineteenth-century magic lantern slides

In visual DH research, classification—dividing a large set of images into several categories—is a common research method. Using the weights of a large model, ImageNet in most cases, as a starting point, researchers can easily train a high-performing classification model using transfer learning. In our first case study, we test the zero-shot capability of CLIP by using it in a binary visual classification task (Smits and Kestemont, 2021). More specifically, we compared the performance of CLIP with a monomodal ResNet-18 CV model and two textual models. The specific binary task concerned recognizing whether a scene depicted on a late-nineteenth-century magic lantern slide was located indoors or outdoors.

The increasing digital availability of lantern slides has stimulated the study of this nineteenth-century ‘mass medium’ (Kember, 2019). The Lucerna Magic Lantern Web Resource, currently hosted by the University of Exeter, is the most important online repository of digitized slides. We collected around 42,000 digitized slides and relevant metadata fields from this online repository. Our dataset contains the URL, filename, title, year of publication, format, people connected to the slide, type of image, dimensions of the slide, materials, production process, persons shown, image content, tags, image location, and location. Using the ‘type of image’ field, we produced a stratified 0.60/0.20/0.20 train, validation, and test set consisting of exterior and interior photographic slides and their captions (Fig. 5).

We used several sets of binary prompts—a set of two words or sentences that are linguistic opposites—to test CLIP’s ability to identify in- and outdoor scenes. As the developers of the model note, choosing the proper prompts is key to achieving adequate performance. This is especially true for our case, as multimodal models can only simulate a binary classification task. After all, seemingly mutually exclusive terms, such as outside and inside, might be similarly unlikely textual descriptions of the same image. In other words, in contrast to the labels of the data used to train a CV model, the prompt ‘outside’ does not necessarily exclude the prompt ‘inside’. Following the application of CLIP to CV benchmarks (Radford *et al.*, 2021), we tested its performance on the stratified test set by using a SoftMax function to normalize the two cosine similarity scores—the result of connecting two prompts per image—into a single probability distribution (Table 1).

We compared CLIP’s performance on this task with a CV model and two textual models that we trained on

our stratified set. For the CV model, we applied a ResNet-18 model, pretrained on the ImageNet dataset. We trained a word unigram model and a character trigram model for the captions. Inspecting the results, Table 1 demonstrates the importance of choosing the right prompts. For example, ‘indoors’ in the ‘outdoors/indoors’ prompt pair achieves a 0.96 accuracy, while ‘outdoor’ in the ‘outdoor/indoor’ pair comes no further than 0.49. On the other hand, ‘exterior’ in the ‘exterior/interior’ pair achieves an accuracy of 0.90, while ‘interior’ remains stuck at 0.71. We experimented with combining high-scoring words from different pairs. However, as Table 1 shows, this did not improve the model’s accuracy, which can be explained by the fact that we applied SoftMax normalization to model’s output for two prompts into a single probability distribution. In other words, the performance of ‘exterior’ depends on the word that is used as its opposite term.

Heeding the advice of CLIP’s developers, we tried to improve the performance by adding contextual information to our prompts. Because we use CLIP to distinguish between two relatively abstract visual concepts, it was unclear which kind of context could lead to better model performance. After trying out several options, we concluded that adding more context did not result in better performance in this specific case. For example, ‘a photograph of an exterior/interior location’ achieved a slightly lower accuracy than the simple ‘exterior/interior’ combination (Table 1).

Overall, the relatively simple CV model outperformed CLIP and the two textual models. It depends on the specific research question whether the difference in performance between CLIP, which does not require labeled data or task-specific training, and the specifically trained monomodal model is problematic. DH researchers need to make a pragmatic decision if improvements in performance warrant the time required to produce labeled data and train models. Our case study shows that even if a multimodal model does perform worse than a purposefully trained visual model, it can still be used to identify patterns in unseen cultural data with high reliability and a significant reduction in training time and thus expenses. As we will argue in the next section, the problem with applying contrastive multimodal models is knowing when good performance is good enough.

4.2 The visual representation of the family in historical children’s literature

The misalignment of the interest of humanities scholars and the abilities of CV models has been a major hurdle for visual DH research. While most scholars are interested in complex and multimodal patterns of representation, CV models can only reliably recognize a limited number of relatively simple visual concepts. Second,



Figure 5. Examples of interior and exterior photographic slides. Reproduced by permission via Lucerna Magic Lantern Web Resource. Left to right: 'Monaco. Monte Carlo Gardens' (item 5047188), slide 24 of 'The Mediterranean' (1887). Image uploaded by Lucerna (2013); 'A Gentleman thought of his Silver and Gold' (item 5073535), slide 8 of 'Our Father's care' (unknown). Image uploaded by Nicholas Hiley (2017)

Table 1. Accuracy of CLIP, a RESNET-18 visual model and two textual models on identifying outdoor/indoor scenes in our stratified set of photographic magic lantern slides.

Accuracy CLIP	Accuracy on exterior	Accuracy on interior	Overall accuracy
Exterior/interior	0.902	0.711	0.807
A photograph of an exterior location/a photograph of an interior location	0.717	0.877	0.797
Outside/inside	0.609	0.931	0.769
Outdoor/indoor	0.498	0.964	0.730
Outdoors/indoors	0.668	0.944	0.806
Exterior/indoor	0.768	0.577	0.673
Street/interior	0.501	0.898	0.699
Accuracy visual model Resnet 18			0.898
Accuracy language models Word unigrams			0.798
Character trigrams			0.777

while humanities scholars regularly work with heterogeneous visual data, most CV models are trained on a single medium, namely digital photographs. In our second case study, we examined CLIP's ability to identify the visual representation of a complex visual concept—the family—in a heterogeneous non-photographic dataset: a collection of around 38,000 realistic, abstract, black-and-white, and color illustrations of Dutch children's books published between 1800 and 1940 (Smits *et al.*, 2022).

Scholars have argued that the interaction between child and family defines children's literature (Stephens, 1992; Alston, 2008). The loss of a beloved family

member, conflicts within a family, or the (dis)harmony between parents and children or brothers and sisters are all common themes. Young readers learn not only about the accepted or desirable social form of a family but also about the multimodal ways in which such a complex social structure is commonly represented. In other words, they learn to recognize what a family is and should be (Stephens, 1992). Scholars of children's literature have noted that while the social make-up of families changed rapidly in the last 50 years, its representations in children's book is marked by stark continuities. Based on the work of Foucault (1969), Alston (2008) describes this as a

tenacious ‘disciplinary discourse’ that reproduces conservative family values.

Most studies of the representation of the family in children’s literature have been based on a close reading of a small number of well-known canonical books. By identifying large-scale patterns of representation, distant reading methodologies can shed light on the ‘slippery’ concept of the family (Wesseling, 2021). While researchers have recently started to apply CV techniques to analyze the pictures of children’s books at scale (Schmideler and Helm, 2021), the wide range of possible (combinations of) family members, family activities, as well as the heterogeneous data—the many different types of illustrations in books for children—make it very hard to train traditional CV models to classify images of the family. For this case study, we examined if CLIP could provide a solution to this problem.

Following the importance of choosing the right prompt, we provided CLIP with three different prompts: ‘a family’, ‘an image of a family’, and ‘an illustration of a family’. We asked the model to return the 5,000 images with the highest cosine similarity score for these prompts, manually annotated the 6,939 images that the model returned, and calculated its precision for every hundredth image (Fig. 6). For the first three prompts, CLIP retrieved, respectively, 1,692, 1,712, and 1,654 correct images. In this case, the model performance increased after adding contextual information (‘an illustration’) to a prompt.

Examining the images that CLIP returned for specific prompts tells us something about the ‘slippery’ representation of the family in children’s literature, but it also reveals how the concept of family is captured in the multimodal model. We can easily spot several recurring themes in the images that the model returned. We see parents and children sitting around a table eating dinner; the mother often cares for a baby or a small child while the father reads the newspaper (Fig. 7). The (eldest) son is often similarly depicted as reading or

doing his homework, while the (eldest) daughter might be knitting, and smaller children are playing on the floor. Interestingly, the characters depicted in these affirmative ‘sex-role stereotypes’ (Stephens, 1992) do not have to be human (Fig. 8). CLIP returned many anthropomorphic images where animals featured as proxies for human families (Stephens, 1992). This ability underlines the power of multimodal models in identifying different visual signals—groups of humans and groups of animals—as belonging to the same conceptual multimodal category.

Next to these patterns in the correctly identified images, we can study common mistakes. The model frequently identified images of girls caring for young(er) children (without parents) as images of the family. Images of girls playing with dolls also belong to this category (Fig. 9). Stephens (1992) already noted that girls are often depicted in the sex-role stereotype of the ‘Big Sister’, which is meant to socialize young female readers into the societal role of being a mother. While CLIP failed to pick up the difference between mother and girl, we could still argue that it had learned the performative and gendered aspects of care in family life. In this sense, CLIP’s mistakes represent the same ‘disciplinary discourse’ about the family as the images that the model identified correctly.

4.3 CLIP as a helping hand for labeling data

In the context of DH research, domain-specific labeled data are often missing. Even though models for image classification, such as ImageNet, contain many different labels, they are often quite generic and lack the specificity that humanities projects require. Fortunately, transfer learning allows us to update existing models using small amounts of training data (Rawat and Wang, 2017; Wevers, 2021). However, this process still requires small amounts of training data and labeling images is a rather time-consuming process. Many

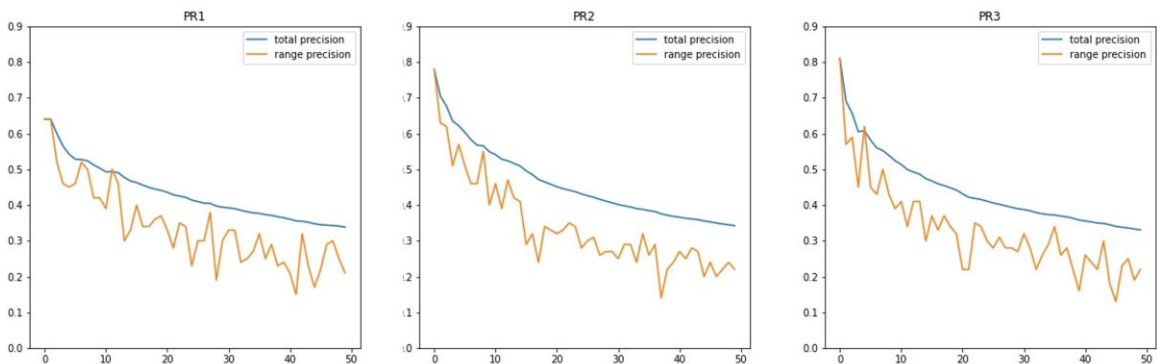


Figure 6. Average precision and range precision per 100 top- n images for the three prompts. From left to right: ‘a family’, ‘an image of a family’, and ‘an illustration of a family’



Figure 7. Recurring themes: the third, fourth, sixth, and seventh highest ranked images for ‘an image of a family’ and the 38.000 images in our corpus (0.324/0.323/0.322/0.321). All images are rights-free via the DBNL, National Library of the Netherlands (left to right): Jan Schenkman, *Torentje torentje bossekruid, of het eerste prentenboek op Moeders schoot* (Amsterdam 1880); anonymous, *Kloentjen, kloentjen garen en anderen* (Haarlem 1888); Dirk Dekker, *Geschiedenis van een rijksdaalder en een cent* (Amsterdam 1874); W.F. Oostveen, *Het sprekend prentenboek* (Rotterdam ca. 1885)



Figure 8. The four highest ranked images of animal families for ‘an image of a family’ and the 38.000 images in our corpus (0.311/0.304/0.293/0.292). All images are rights-free via the DBNL, National Library of the Netherlands (left to right): Mattheus van Heijningen Bosch, *Jan en zijn zusje, of eerste leeslesjes* (Groningen, 1818); anonymous, *Leerzaam allerlei voor de lieve kleinen* (Rotterdam 1857); Oom Anton, *De prentjes van Oom Anton met versjes erbij* (Amsterdam 1920–30); anonymous, *Nieuwe dieren gallery voor kinderen* (Mainz 1870–80)



Figure 9. The four highest ranked images of girls caring for younger siblings and/or dolls for ‘an image of a family’ and the 38.000 images in our corpus (0.318/0.313/0.309/0.306). All images are rights-free via the DBNL, National Library of the Netherlands (left to right): J.J.A. Goeverneur, *Fabelen en gedichtjes* (Leeuwarden 1873); anonymous, *De koningin der poppen* (Amsterdam 1867); Jacob van Lennep, J.J.L. ten Kate en S.J. van den Bergh, *De nachtegaal en het lijstertje* (Leiden 1854); Anna Sutorius, *Pannekoeken bakken* (Den Haag 1929)

projects certainly lack the time and funds to produce high-quality labeled data.

In the context of the Photographic Memory project, we use CLIP to generate labels for unlabeled training data (Wevers *et al.*, 2022). Using a predefined list of labels, we apply CLIP’s zero-shot capability to suggest labels for images. While CLIP is trained on millions of

contemporary image–text combinations extracted from the Internet, it can still provide good results in zero-shot classification tasks applied to images the model was never exposed to during training. However, it is unknown how this performance varies for specific visual source materials or labels. Therefore, we validated the generated labels using crowd workers and domain experts.

Rather than having crowd workers annotate a large share of the data, we used smaller samples of images that were enriched with labels provided by CLIP and an existing model (PLACES-365). Initially, we relied on an existing model (PLACES-365) to label our images. Yet, we discovered that when combining this model with CLIP, rather than just relying on CLIP, and averaging the predictions of the labels, we acquired better results. In an iterative fashion, crowd workers checked these labels. The correct labels were subsequently used as training data to improve the PLACES-365 model using transfer learning as well as to train a classifier on top of the existing CLIP model.

After going through three stages, the PLACES-365 as well the CLIP model reached a much higher accuracy and we were able to provide many of the images with correct labels.² We decided to not only rely on CLIP but also train an additional model for the specific prediction task. This was done for two reasons. First, CLIP, as mentioned above, is a rather heavy solution for a somewhat simple prediction task. To improve the re-usability of the project's model, we decided to also provide a more lightweight model specifically adapted to the task of scene detection. Second, it is not clear what biases exist in the CLIP model as the training data are unknown. To improve transparency, we will also publish the updated PLACES-365 model. For the latter model, it is known what training data were used as well as the data that we used to update the model.

Overall, this case study shows how CLIP can function as a helping hand in generating labeled training data in low-resource contexts or contexts with smaller amounts of training data. This iterative training process can still be extended and improved, but it nonetheless shows the potential use of CLIP in a wide array of domains and applications.

5 Multimodal bias and DH research

As the case studies showed, applying multimodal models to visual heritage collections opens up many exciting new pathways for DH research. However, the introduction of these models also comes with a specific set of new methodological challenges. This section argues that researchers must take the bias of these models into account: the parts of the visual word that they make visible and obscure. Just like other CV techniques, multimodal models project a specific part of the contemporary world—the data on which they were trained—on to (historical) sources.

To shed light on the visual bias of CV models, researchers have mainly pointed to the collection, annotation, and organization of the large training sets (Jo and Gebru, 2020). In a pathbreaking study, Crawford and Paglen (2019) described how the three most

important layers of these datasets—the taxonomy of the classes, the individual classes, and the individually labeled images—led to biased predictions. Although multimodal models require no taxonomy or classes, Birhane *et al.* (2021) demonstrate how their training data influences their predictions. This kind of critical analysis of training data is made difficult by the fact that CLIP and most other multimodal models do not publicly release the data on which they were trained. Efforts have been made to replicate the dataset used to train CLIP, resulting in the LAION-400M WebImageText dataset (Schuhmann *et al.*, 2021). Birhane *et al.* (2021) use this dataset as a proxy to flag the large amounts of ‘problematic’ and NSFW (not suitable for work) content used to train CLIP, which consists of ‘explicit images and text pairs of rape, pornography, malign stereotypes, [and] racists and ethnic slurs.’ They also show that ALT text, the textual input of the HTML element used in CLIP’s training data, frequently describes images in stereotypical and offensive ways. Because a lot of image–text pairs on the Internet are pornographic in nature, relatively ‘benign’ textual descriptions are connected to explicit visual concepts. In the LAION-400M dataset, words like ‘mom’, ‘nun’, ‘daddy’, and ‘schoolgirl’ frequently appear with sexual visual content. Because the multimodal understanding of models like CLIP is dependent on its training data, the visual concepts it learns are ‘more representative’, as CLIP’s model card puts it (CLIP Model Card, 2020/2022), of the worldview of persons that are most connected to the Internet: young, male, and English-speaking users that live in developed nations. To address these problems, the developers of LAION-2B dataset, a follow-up to LAION-400M, flagged and watermarked NSFW content (Schuhmann *et al.*, 2021, 2022). This allows researchers to examine how this content relates to specific prompts.³ In theory, these images can also be filtered out from the dataset when training a model.

Multimodal models are not only biased toward the worldview of the most avid Internet users, the way in which they see the world is also tied to a specific historical period. We previously argued that scholars need to take ‘historical bias’ into account when they apply CV models to historical collections (Smits and Wevers, 2021). Trained on modern photographs, CV models cannot help but look for non-relevant modern categories, such as ‘parking meter’ or ‘computer mouse’, in historical images. The training data of multimodal models are derived from the Common Crawl public web archive, which, since 2008, publishes snapshots of the Internet. While the Internet contains many images, such as digitized photographs and paintings, that were made before 2008, the ALT text descriptions of these pictures all interpret them within a specific time frame.

Just like CV models, contrastive multimodal models project a particular semiotic mode—the image–text combinations of the HTML ALT text element—of a particular time—the Internet between 2008 and 2021—and a specific social world—the one created by young male users in western countries—on to new image–text combinations in the future and, if applied by historians, the past.

How does the historical bias of multimodal models manifest itself in DH research? After we presented our work on the retrieval of images of the family in children’s books, several colleagues asked if the model would be able to retrieve images of lgbtqi+ parents. Being trained on contemporary material, CLIP will, in fact, be very likely to ‘wrongly’ identify images of two women or two men and a child as an image of a family. After all, it is improbable that the nineteenth-century source material would include visual representations of this type of family. In this case, the model ‘identifies’ a contemporary concept in historical sources where we know it is not present. On the opposite side of this coin, a multimodal model might fail to identify concepts that were not present in its contemporary training material. For example, we know that in many nineteenth-century bourgeois households, nurses and maids looked after children. However, CLIP probably is unable to recognize the difference between a mother and a nurse because this latter category is not present in its training material.

How should DH researchers deal with the bias of multimodal models? First of all, we need more studies like [Birhane et al. \(2021\)](#) that look critically at multimodal models and the data on which they were trained to better understand the specific shape of their (historical) bias. This will allow us to better assess the effects of these biases in research. Second, scholars need to reflect on how these models’ (historical) bias might interfere with or influence their specific research question. Did the visual concept that we are trying to identify exist (in the same form) in the time of our sources/when the training data for the multimodal model were collected? In many ways, this is a normal critical historical thinking. However, it does require a basic understanding of how multimodal machine learning works. Such an understanding might help in constructing prompts or ensembles of prompts that can mitigate bias in models. Third, depending on the research question and our domain knowledge, we can also, as our third case study shows, fine-tune CLIP or add a classifier to the embeddings that CLIP produces to create a model that is better attuned to our data and our question. Moreover, recent efforts in which few-shot learners are added to existing models to increase their performance on specific tasks offer promising possibilities for humanities research ([Tsimpoukelli et al., 2021](#)).

6 Conclusion

Will multimodal models cause a multimodal turn in DH research? On a methodological level, this seems probable. As a result of their specific task (connecting text and images) and the impressive amounts of data on which they were trained, multimodal models can be applied to a wide range of text-to-image, image-to-text, and image-to-image retrieval and classification tasks. Consequently, researchers can easily use sophisticated deep-learning techniques without having to label data or train models themselves. Like OCR for textual archives, multimodal models provide a radically new kind of bottom-up access to visual collections. Researchers will not be the only ones to benefit from this new technique. As a result of their zero-shot capability, many different types of users—librarians, researchers, and the larger public—can use of the same embeddings without additional training to perform a wide range of different tasks.

We expect that a possible multimodal turn in DH will not only be practical in nature. Studying images, especially at scale, has always been a daunting task. Even if images are described through metadata, the images in digital collections always hold more information and possible meanings than could fit on the back of a punch card or that we might reasonably expect a librarian to enter into an archiving system. As [Sontag \(1977\)](#) noted, images need text to anchor them in the many meanings that they can convey or, maybe more to the point, we, as viewers of images, need texts to understand and contextualize them. Transferred to the situation of DH scholars, we argue that researchers need to connect images to texts to study and understand them. This article shows that contrastive multimodal models can help scholars to connect images and texts and examine how these two elements interact with each other on an unprecedented scale.

Author contributions

Thomas Smits (Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, Writing—review and editing), Melvin Wevers (Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, Writing—review and editing).

Notes

1. <https://kiosk-dot-codelabs-site.appspot.com/codelabs/tensor-flow-for-poets-2-ios/index.html?index=..%2F..index#0>
2. For the scene detection task Places-365 reached an accuracy of 0.58 and a top-5 accuracy of 0.86. CLIP reached an accuracy of 0.64 and a top-5 accuracy of 0.91. For detecting

whether an image was taken indoors or outdoors, we see that places-365 reached a 0.93 accuracy and CLIP 0.94. For more information see the model cards: <https://github.com/melvinwevers/HisVis2/tree/main/docs>

- The developers also released this tool to explore the training data: <https://knn5.laion.ai>

References

- Alston, A. (2008). *The Family in English Children's Literature*. New York: Routledge.
- Arnold, T. and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, 34(Supplement_1): i3–16. <https://doi.org/10.1093/lc/fqz013>
- Barthes, R. (1961). Le message photographique. *Communications*, 1(1): 127–38. <https://doi.org/10.3406/comm.1961.921>
- Bateman, J. A. (2014). *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. Routledge.
- Bingham, A. (2010). The digitization of newspaper archives: opportunities and challenges for historians. *Twentieth Century British History*, 21(2): 225–31.
- Birhane, A., Prabh, V. U., and Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. ArXiv:2110.01963 [Cs]. <http://arxiv.org/abs/2110.01963>
- Bod, R. (2013). Who's afraid of Patterns?: The Particular versus the Universal and the Meaning of Humanities 3.0. *BMGN - Low Countries Historical Review*, 128(4): 171–80. <https://doi.org/10.18352/bmgn-lchr.9351>
- Champion, E. M. (2017). Digital humanities is text heavy, visualization light, and simulation poor. *Digital Scholarship in the Humanities*, 32(Suppl_1): i25–32. <https://doi.org/10.1093/lc/fqw053>
- CLIP Model Card. (2020/2022). [Jupyter Notebook]. OpenAI. <https://github.com/openai/CLIP/blob/d50d76daa670286dd6caf3bcd80b5e4823fc8e1/model-card.md> (Original work published 2020).
- Crawford, K. and Paglen, T. (2019, September 19). *Excavating AI: The Politics of Training Sets for Machine Learning*. <https://www.excavating.ai>
- Desai, K. and Johnson, J. (2021). *VirTex: Learning Visual Representations from Textual Annotations*. pp. 11162–73. https://openaccess.thecvf.com/content/CVPR2021/html/Desai_VirTex_Learning_Visual_Representations_From_Textual_Annotations_CVPR_2021_paper.html
- Duhaime, D. and Leonard, P. (2017/2020). *PixPlot* [JavaScript]. Yale Digital Humanities Lab. <https://github.com/YaleDHLab/pix-plot> (Original work published 2017).
- Foucault, M. (1969). *L'archéologie du Savoir*. Gallimard.
- Hiipala, T. (2021). Distant viewing and multimodality theory: prospects and challenges. *Multimodality & Society*, 1(2): 134–52. 26349795211007096. <https://doi.org/10.1177/26349795211007094>
- Jain, T., Lennan, C., and Train, D. (2019/2020). *IdealolImagededup* [Python]. <https://github.com/idealol/imagededup> (Original work published 2019).
- Jia, C., Yang, Y., Xia, Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR, virtual, pp. 4904–16. <https://proceedings.mlr.press/v139/jia21b.html>
- Jo, E. S. and Gebru, T. (2020). Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, pp. 306–16. <https://doi.org/10.1145/3351095.3372829>
- Kember, J. (2019). The magic lantern: open medium. *Early Popular Visual Culture*, 17(1): 1–8. <https://doi.org/10.1080/17460654.2019.1640605>
- Kuznetsova, A., Rom, H., Alldrin, N., et al. (2020). The open images dataset v4. *International Journal of Computer Vision*, 128(7): 1956–81.
- Lectaru, K. (2008). Mass book digitization: The deeper story of Google Books and the Open Content Alliance. *First Monday*. <https://doi.org/10.5210/fm.v13i10.2101>
- Lin, T.-Y., Maire, M., Belongie, S., et al. (2014). Microsoft coco: common objects in context. In *European Conference on Computer Vision*, pp. 740–55.
- Manovich, L. (2020). *Cultural Analytics*. The MIT Press.
- Mitchell, W. J. T. (2005). There are no visual media. *Journal of Visual Culture*, 4(2): 257–66.
- Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 1: 54–68.
- Moretti, F. (2015). *Distant Reading*. Verso.
- Nicholson, B. (2013). The digital turn. *Media History*, 19(1): 59–73.
- Paul, G. (2016). *Das visuelle Zeitalter: Punkt und Pixel*. Wallstein Verlag.
- Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, PMLR, virtual, pp. 8748–63.
- Rawat, W. and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Computation*, 29(9): 2352–449. https://doi.org/10.1162/neco_a_00990
- Reese, H. and Heath, N. (2016, December 16). *Inside Amazon's Clickworker Platform: How Half a Million People Are Being Paid Pennies to Train AI*. TechRepublic. <https://www.techrepublic.com/article/inside-amazons-clickworker-platform-how-half-a-million-people-are-training-ai-for-pennies-per-task/>
- Russakovsky, O., Deng, J., Su, H., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–52.
- Schmideler, S. and Helm, W. (eds) (2021). *BildWissen—KinderBuch*. J.B. Metzler. <https://doi.org/10.1007/978-3-476-05758-7>
- Schuhmann, C., Beaumont, R., Vencu, R., et al. (2022). LAION-5B: an open large-scale dataset for training next generation image-text models (arXiv:2210.08402). arXiv. <https://doi.org/10.48550/arXiv.2210.08402>
- Schuhmann, C., Vencu, R., Beaumont, R., et al. (2021). LAION-400M: open dataset of CLIP-filtered 400 million image-text pairs (arXiv:2111.02114). arXiv. <https://doi.org/10.48550/arXiv.2111.02114>
- Seguin, B., di Leonardo, I., and Kaplan, F. (2017, August 11). *Tracking Transmission of Details in Paintings*. Montreal: Digital Humanities.

- Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12): 1349–80. <https://doi.org/10.1109/34.895972>
- Smits, T., van der Eecken, P., and Joosen, V. (2022, May 9). *Using CLIP to Extract and Analyze Images of the Family in 3,000 Dutch-Language Children's Books, 1800–1940*. DH Benelux 2022—ReMIX: Creation and Alteration in DH (hybrid), Zenodo. <https://doi.org/10.5281/zenodo.6530429>
- Smits, T. and Kestemont, M. (2021). Towards multimodal computational humanities. Using CLIP to analyze late-nineteenth century magic lantern slides. In Ehrmann, M., Karsdorp, F., Wevers, M., et al. (eds), *Proceedings of the Conference on Computational Humanities Research 2021*, Vol. 2989, pp. 149–58. CEUR. http://ceur-ws.org/Vol-2989/#short_paper23
- Smits, T. and Ros, R. (2021). Distant reading 940,000 online circulations of 26 iconic photographs. *New Media & Society*, 14614448211049460. <https://doi.org/10.1177/14614448211049459>
- Smits, T. and Wevers, M. (2021). The agency of computer vision models as optical instruments. *Visual Communication*, 21(2): 329–49. <https://doi.org/10.1177/1470357221992097>
- Sontag, S. (1977). *On Photography*. Farrar, Straus and Giroux.
- Stephens, J. (1992). *Language and Ideology in Children's Fiction*. Longman.
- Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S. M. A., Vinyals, O., and Hill, F. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–12. <https://proceedings.neurips.cc/paper/2021/hash/01b7575c38dac42f3cfb7d500438b875-Abstract.html>
- van Noord, N. (2022). A survey of computational methods for iconic image analysis. *Digital Scholarship in the Humanities*, 37(4): 1316–1338. <https://doi.org/10.1093/llc/fqac003>
- Wesseling, L. (2021). Family. In Nel, P., Paul, L., and Christensen, N. (eds), *Keywords for Children's Literature*, Vol. 9, pp. 74–77. NYU Press.
- Wevers, M. (2021). Scene detection in de Boer historical photo collection. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. Vienna: ICAART.
- Wevers, M. and Lonij, J. (2017, October 15). Siamese. *KB Lab*. <http://lab.kb.nl/tool/siamese>
- Wevers, M. and Smits, T. (2020). The visual digital turn. Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35(1): 194–207. <https://doi.org/10.1093/llc/fqy085>
- Wevers, M., Vriend, N., and de Bruin, A. (2022). What to do with 2.000.000 historical press photos? The challenges and opportunities of applying a scene detection algorithm to a digitised press photo collection. *TMG Journal for Media History*, 25(1): Article 1. <https://doi.org/10.18146/tmg.815>
- Wijfjes, H. (2017). Digital humanities and media history: a challenge for historical newspaper research. *TMG Journal for Media History*, 20(1): Article 1. <https://doi.org/10.18146/2213-7653.2017.277>