

Visual Navigation of Digital Libraries: Retrieval and Classification of Images in the National Library of Norway's Digitised Book Collection

Marie Roald^{1,*}, Magnus Breder Birkenes¹ and Lars Gunnarsønn Bagøien Johnsen¹

¹Research and Special Collections, The National Library of Norway, Norway

Abstract

Digital tools for text analysis have long been essential for the searchability and accessibility of digitised library collections. Recent computer vision advances have introduced similar capabilities for visual materials, with deep learning-based embeddings showing promise for analysing visual heritage. Given that many books feature visuals in addition to text, taking advantage of these breakthroughs is critical to making library collections open and accessible. In this work, we present a proof-of-concept image search application for exploring images in the National Library of Norway's pre-1900 books, comparing Vision Transformer (ViT), Contrastive Language-Image Pre-training (CLIP), and Sigmoid loss for Language-Image Pre-training (SigLIP) embeddings for image retrieval and classification. Our results show that the application performs well for exact image retrieval, with SigLIP embeddings slightly outperforming CLIP and ViT in both retrieval and classification tasks. Additionally, SigLIP-based image classification can aid in cleaning image datasets from a digitisation pipeline.

Keywords

image retrieval, computer vision, embeddings, vector search

1. Introduction

With the goal of preserving and disseminating Norwegian cultural heritage, the National Library of Norway (NLN) began digitising its collection in 2006. This collection, acquired per the Norwegian Legal Deposit Act¹, spans various materials, including books, newspapers, journals, posters, radio, movies and more [4]. Almost all books and most newspapers have already been digitised, barring a few exceptions, and the current focus is on processing newspapers, journals, and non-text-based media [4]. However, digitisation alone is insufficient to make cultural heritage available; it is also necessary to ensure that the digitised content is easy to view and access is not overly restricted. Thus, the *Bokhylla* agreement grants regulated access [11], and the online library *Nettbiblioteket* lets users view collections with an International Image Interoperability Framework (IIIF) [23] based viewer and perform full-text searches using Elasticsearch. Finally, NLN offers limited access to the textual content through NB DH-LAB


CHR24: 5th Conference on Computational Humanities Research, December 04–06, 2024, Aarhus, Denmark

*Corresponding author.

✉ marie.roald@nb.no (M. Roald); magnus.birkenes@nb.no (M. B. Birkenes); lars.johnsen@nb.no

(L. G. B. Johnsen)

ORCID 0000-0002-9571-8829 (M. Roald)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://lovdata.no/dokument/NL/lov/1989-06-09-32>

[4] and corresponding webapps² which provides tools based on text aggregates (e.g. n-grams, collocations and concordances) to facilitate automated and reproducible analysis of the text.

Currently, these tools have largely been based on text extracted from Analysed Layout and Text Object-Extensible Markup Language (ALTO-XML) files³ generated by optical character recognition (OCR) models during digitisation [5]. However, the output XML also contains coordinates for graphical elements. These graphical elements represent non-textual elements in the books, e.g. illustrations or decorations. While such elements are an important part of the books, they have been cumbersome to explore, requiring manual inspection. Therefore, an essential missing step for making NLN's digitised collection more accessible is making these graphical elements easier to explore and analyse.

An approach to make such elements explorable, is creating tools for image search, either in the form of exact image retrieval (i.e. recovering a specific image) or semantic image retrieval (i.e. recovering an image with similar contents) or both. While text-based search engines are commonplace, image search is more complicated [16, 28]. Early methods matched images using surrounding text [16], but this approach demands high-quality textual descriptions, which can be lacking. Alternatively, exact image retrieval traditionally relies on handcrafted image features for comparison [16, 28]. Handcrafting such features can be challenging, and typically form a dense vector, which can hinder efficient lookups.

However, recent technological advancements have simplified the implementation of image search engines. Various tools now implement efficient search indices for dense vectors, such as the hierarchical navigable small worlds (HNSW) index [12]. Moreover convolutional neural networks (CNNs) and vision transformers (ViTs) have alleviated the need for handcrafted image features for computer vision [8, 7]. Furthermore, there has been an influx of multi-modal models, like Contrastive Language-Image Pre-training (CLIP) [19] and Sigmoid Loss for Language Image Pre-Training (SigLIP) [27]. The recent advances in computer vision and proliferation of advanced pre-trained computer vision models has empowered the development of new research and tools for exploring and analysing image-based data in the digital humanities [2, 25, 21, 9, 22, 20].

Previous work on machine learning-driven computer vision-based image search tools for digital humanities mainly focuses on cleanly digitised materials such as collections of videos, photographs, lantern slides and medieval illuminations [2, 21, 22, 17]. However, there is limited work applying such tools to images extracted from the output of automatic layout detection of scanned media, e.g. books and newspapers. Such image collections pose unique challenges. First, the magnitude of data is often larger than for collections of photographs. Second, such data can contain artefacts not found in cleanly digitised materials. For example, detected bounding boxes might be inaccurate. False positives can occur, where the automatic layout detection mistakenly marks, e.g. tables or blank pages, as graphical elements. Avoiding such artefacts can be infeasible, as redoing layout analysis for a collection of sizeable magnitude can be cost-prohibitive and not guaranteed to succeed. Therefore, a natural next step is exploring machine learning-based image retrieval in the context of NLN's collection of scanned automatically processed media.

²<https://www.nb.no/dh-lab/apper/>

³<https://www.loc.gov/standards/alto/>

This short paper details ongoing work on these challenges, with three primary contributions:

1. Developing a proof-of-concept image search application for NLN’s pre-1900 books.
2. Comparing modern image embeddings for image retrieval in NLN’s digitised books.
3. Evaluating pre-trained models for fine-tuned classification of image categories.

2. Background and related work

Two traditional approaches for image retrieval are context-based full-text search – querying the images’ textual context – and hashing-based approaches for exact image retrieval. The former typically works by using an inverted index to efficiently retrieve relevant images via e.g. term frequency-inverse document frequency (TF-IDF) weighting [24], before potentially re-ranking them based on image features [16]. The hashing-based alternative works by computing a compact hash, or “fingerprint”, that can be used for efficient exact image retrieval [6].

More recent image retrieval approaches compute image similarities using deep learning-based image classification models such as ViTs [7] or CNNs [8]. These models first transform an image into an embedding, which is used as input for a logistic regression model. The key insight in using these models for image retrieval is that we can compute image similarities by comparing the embeddings, e.g. with the cosine similarity.

However, by using classification models, we assume that embeddings learned by training on image-label combinations are informative enough to group images semantically, which can hinder generalisation to out-of-sample images [15]. Another approach is multimodal models like CLIP and SigLIP. In short, these models work by combining an image transformer and a text transformer to compute image and text embeddings – aligning them to ensure strong cosine similarity for matching pairs. This approach has been successfully applied to e.g. image retrieval and zero-shot classification [19], and generalise better to out-of-sample images [19, 15].

During CLIP and SigLIP training, models receive shuffled image-caption pairs and compute probabilities for matches. Such training demands extensive data and computational resources. To circumvent this, it is common to use pre-trained models and the popularity of model repositories, like Huggingface Hub [26] and Torch Hub [1], has made using models trained on massive datasets accessible.

While methods for efficient sparse vector queries have existed for decades [10], querying based on image embeddings requires dense vector queries, which is still a research topic. However, the recently proposed HNSW-index for approximate nearest neighbour search [12] has gained traction for accuracy and efficiency. The index consists of a hierarchy of navigable small world graphs [13], each built from different data subsets, and querying consists of iteratively traversing the hierarchy, enabling efficient navigation through large datasets.

Applying modern computer vision to problems in digital humanities has recently gained traction. The term *distant viewing* is introduced in [2], which demonstrates how computer vision methods for clustering and object detection can be applied to image- and video-data. Building on this, [25] shows how CNN-based semantic image retrieval can be used to explore trends in newspaper advertisements and illustrations extracted from Delpher – a digitised materials search engine by the Dutch national library. Moreover, [17] demonstrate how a

combination of monomodal image- and language-models can be used to combine and enrich two manually annotated collections of medieval illuminations and [21, 22] shows how a CLIP model can be used to explore and label magic lantern slides efficiently and that it can struggle with zero-shot classification of old illustrations. Using CLIP embeddings, [20] clusters news videos and employs a graph-based approach for efficient exploration. Machine learning-driven image retrieval tools for libraries and museums, like Maken⁴, Bildsök⁵ and Nasjonalmuseet Beta⁶ have also emerged. These previous works highlight computer vision’s potential in digital humanities, and thus, evaluating and comparing such models in the context of NLN’s digitised book collection is a relevant next step.

3. Methods

3.1. Extracting images

To search the images, they must first be extracted from the digitised book collection. During NLN’s digitisation, books are scanned and processed through a pipeline including layout detection and OCR, producing ALTO-XML files⁷ named after Uniform Resource Names (URNs). These files contain page information, describing the page in terms of four block types: TextBlock, Illustration, GraphicalElement and CompositeBlock (blocks containing other blocks)⁸. In the ALTO-XML files parsed for this work, all illustrations and graphical elements are tagged as GraphicalElement. Parsing these files, we extracted the page URN, coordinates, and size for each graphical element in addition to the textual context of each image in the digitised books. For this work, we processed pre-1900 books, creating a sufficiently large, yet manageable subset for testing.

For each graphical element, we used NLN’s IIIF API⁹ to download images from URLs following the format in Table 1, discarding images with aspect-ratio ≥ 50 . By integrating ALTO-XML files with the IIIF endpoint – both technologies already utilised by NLN – we obtained images from digitised Norwegian books before 1900.

3.2. Creating the vector search application

We computed image embeddings using Huggingface Transformers [26] with three models: ViT (google/vit-base-patch16-224¹⁰), CLIP (openai/clip-vit-base-patch32¹¹) and SigLIP (google/siglip-base-patch16-256-multilingual¹²). Each pre-trained model’s preprocessing pipeline involved resizing images to the input shapes (224 for ViT and CLIP, and 256 for SigLIP) and scaling the pixel values. For ViT and SigLIP, images were resized to 224 × 224 and

⁴<https://www.nb.no/maken/>

⁵<https://lab.kb.se/bildsok/>

⁶<https://beta.nasjonalmuseet.no/collection/>

⁷<https://digitalpreservation-blog.nb.no/docs/formats/preferred-formats-en/>

⁸<https://www.loc.gov/standards/alto/techcenter/layout.html>

⁹<https://iiif.io/api/image/2.0/>

¹⁰Commit hash: 3f49326eb077187dfe1c2a2bb15fbd74e6ab91e3

¹¹Commit hash: 3d74acf9a28c67741b2f4f2ea7635f0aaf6f0268

¹²Commit hash: a66c5982c8c396206b96060e2bf837d6731a326f

Table 1
The IIIF URL format.

	Description	Example
	Scheme	https://
	Prefix	www.nb.no/services/image/resolver/
	Identifier (URN)	URN:NBN:no-nb_digibok_2009070210001_0618/
Region (left, top, width, height)		430, 432, 2195, 2348/
	Size (width, height)	274, 294/
	Rotation (degrees)	0/
Filename (filename.filetype)		default.jpg
	Full URL	https://www.nb.no/services/image/resolver/URN:NBN:no-nb_digibok_2009070210001_0618/430,432,2195,2348/274,294/0/default.jpg

256 × 256 pixels, altering the aspect ratio. CLIP resized the smallest dimension to 224, preserving the aspect ratio, then center-cropped to 224 × 224 pixels. Next, we used the corresponding image transformer and obtained embeddings of sizes 768 (ViT and SigLIP) and 512 (CLIP).

After computing embeddings, we ingested them into a Qdrant database and used FastAPI to create an application programming interface (API) for efficient querying by images, embedding vectors, image IDs, or context-based text search. Qdrant supports fast K-nearest neighbour search for both dense and sparse vectors. For image-based queries, we used a cosine similarity-based HNSW index, and for context-based full-text queries, we used a dot-product-based inverted index for TF-IDF (details in supplement on GitHub¹³). We used default parameters for all search indices. The vector database and the API are hosted on-premise, exposing only the API to the Internet. The application also includes a frontend, implemented using Flask and HTMX, hosted using Google Cloud Run with 512 MiB RAM and one vCPU.

3.3. Classifying based on embedding vectors

As the graphical elements stem from NLN’s digitisation process, many segmentation anomalies are also tagged as graphical elements. Common examples are blank pages, parts of tables, and text. To estimate the fraction of such regions, we used HumanSignal Label Studio and manually labelled a dataset containing 2000 images as either *Blank page*, *Segmentation anomaly*, *Illustration or photograph*, *Musical notation*, *Map*, *Mathematical chart* or *Graphical element* (e.g. initial, decorative border, etc.).

After labelling the data, we fitted regularised logistic regression models (using scikit-learn v1.5.0 [18]) to classify images based on their embedding vectors. This can be interpreted as a form of transfer learning, fine-tuning the last layer of the transformer model. The embedding vector type (i.e. ViT, CLIP or SigLIP) and the complexity parameter (inverse ridge parameter) were selected using nested cross-validation with 20 outer folds and ten inner folds. Models were selected based on a micro-averaged F1-score (the harmonic mean of micro-averaged precision and sensitivity). We selected the complexity parameter from ten logarithmically spaced values

¹³<https://github.com/Sprakbanken/CHR24-image-retrieval>

between 10^{-4} and 10^4 . Finally, we computed the confusion matrix in the outer cross-validation loop (the evaluation loop). The supplement describes the overall cross-validation algorithm in Algorithms 1 and 2.

3.4. Evaluating searches

To evaluate the search, we first manually inspected some example queries before performing a systematic evaluation on exact image retrieval. To simulate exact image retrieval scenarios, we selected the 684 images labelled as *Illustration or photograph*, *Map* or *Mathematical chart* as target images, and applied random cropping ($\leq 15\%$, independently on all sides), rotation ($\pm 0 - 10^\circ$) and scaling ($\pm 0 - 20\%$, independently for width and height). Then, querying the database with these transformed images, we evaluated the Top N accuracy measuring whether our application retrieved the target image in the first result (Top 1), first row (Top 5), first two rows (Top 10) or results at all (Top 50).

4. Results

Figure 1 shows screenshots from the application¹⁴ for image searches using full-text (Fig. 1a) or image similarity (Figs. 1b to 1d). Table 2 shows image-based query results with four different images. For the first row, the query exists in the collection, and all models recover it as the top result. Similarly, for the second row, all models return nautical results, and CLIP is the only model that does not return illustrations with lighthouses. Finally, the third and fourth rows show examples of querying with images outside of the collection, where we see that the returned images are content-wise similar. The fourth row demonstrates an example where CLIP embedding vectors fail, leading to irrelevant results. Furthermore, the exact image retrieval experiments demonstrate that our application can recover queried transformed images. As demonstrated in Table 3, SigLIP performed slightly better than ViT and CLIP and retrieved 94% of the target images in the first two rows of the search and 97% in all ten displayed rows. See GitHub for code and details.

The manual image labelling¹⁵ showed that 349/2000 (17%) of the graphical elements were blank pages and 524/2000 (26%) were segmentation anomalies (e.g. tables, text, etc.) – for complete label distribution, see Fig. 2. Moreover, the logistic regression model performs well, obtaining a cross-validated F1 score of 96% ($\sigma = 5.1\%$). From the cross-validated confusion matrix, we see that only 66/1127 ($< 6\%$) of all graphical elements were incorrectly classified as either blank pages or segmentation errors, with a marked amount of incorrect classifications being from the “Graphical element” class. We also observed that the SigLIP embeddings were selected in all 20 outer cross-validation folds, indicating their superiority for this classification task compared to ViT and CLIP. Fig. 2 also shows the estimated class distribution on the full dataset.

¹⁴<https://dh.nb.no/run/bildesok/>

¹⁵The labels and analysis code are available on GitHub

5. Discussion and conclusion

These promising results demonstrate that pre-trained computer vision models provide meaningful embeddings. This is notable as our data consists of pre-1900 book images and differs vastly from the training set of such models, which are typically scraped from the internet. Furthermore, the results indicate that SigLIP embeddings slightly outperforms CLIP and ViT for all tasks – even for image classification, which ViT was trained for – in line with prior results showing that multimodal models are more robust to out-of-sample data [15].

While all models perform well for retrieval, CLIP sometimes struggled, particularly if the object of interest was off-centre. In such cases, the object is cropped out during preprocessing and matches will be based on the remaining image. Furthermore, the application performs well for exact image retrieval, even with up to 30 % cropping in both directions and up to $\pm 10^\circ$ rotation. These results are promising, but more work is still needed to evaluate performance for other degradations (e.g. simulated print and scanning artefacts). Finally, the encouraging image classification results indicate advantages of adding this methodology to the data ingestion pipeline. Filtering out irrelevant elements can save up to 40 % storage and improve the search results.

In conclusion, we found that by combining tagged graphical elements of the book digitisation process, NLN’s IIF endpoint and recent advances in artificial intelligence, we can create an efficient image search application that facilitates exploring the library’s collection in a new way.

6. Future work

As the current prototype image-search app only supports books pre-1900, a natural extension is including illustration objects from all NLN’s digitised books and newspapers. Moreover, as one use case we consider is exact image retrieval, an obvious next step is more thorough analysis of the the application’s accuracy on this task, e.g. using additional evaluation measurements for recall, and including domain-specific degradation (e.g. simulated halftone and scanning artefacts). Another avenue for future work is comparing deep learning-based similarity measures with simpler, less computation- and storage-intensive approaches like hashing-based methods. Additionally, we want to make the software more adaptable, ultimately creating open-source infrastructure to further these methods’ accessibility for other ALTO-XML and IIF collections.

Future work should explore the embeddings further, e.g. using CLIP and SigLIP for text-based image retrieval. Additionally, performance could improve by fine-tuning the embeddings on domain-relevant data. Moreover, we have so far only used the embeddings for image retrieval and classification. Using the embeddings as the base to discover clusters, automatically tag the images or create image descriptions are, therefore, interesting potential steps. Another important direction is digging deeper into what the models consider ”similar” through visualisations and empirical experiments. Finally, because deep learning-based embeddings are trained on datasets with known biases [3, 22, 14], examining biases in these embeddings is crucial.

References

- [1] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. K. Luk, B. Maher, Y. Pan, C. Puhersch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, S. Zhang, M. Suo, P. Tillet, X. Zhao, E. Wang, K. Zhou, R. Zou, X. Wang, A. Mathews, W. Wen, G. Chanan, P. Wu, and S. Chintala. “PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation”. In: *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. La Jolla, CA, USA, 2024, pp. 929–947. DOI: 10.1145/3620665.3640366.
- [2] T. Arnold and L. Tilton. “Distant Viewing: Analyzing Large Visual Corpora”. In: *Digital Scholarship in the Humanities* 34.Supplement_1 (2019), pp. i3–i16. DOI: 10.1093/llc/fqz013.
- [3] A. Birhane, V. U. Prabhu, and E. Kahembwe. *Multimodal datasets: misogyny, pornography, and malignant stereotypes*. <https://arxiv.org/abs/2110.01963>. 2021. arXiv: 2110.01963.
- [4] M. B. Birkenes, L. Johnsen, and A. Kåsen. “NB DH-LAB: A Corpus Infrastructure for Social Sciences and Humanities Computing”. In: *CLARIN Annual Conference Proceedings 2023*. Leuven, Belgium, 2023, pp. 30–34.
- [5] M. B. Birkenes, L. G. Johnsen, A. M. Lindstad, and J. Ostad. “From Digital Library to N-Grams: NB N-gram”. In: *Proceedings of the 20th Nordic Conference of Computational Linguistics*. Vilnius, Lithuania, 2015, pp. 293–295.
- [6] R. Biswas and P. Blanco-Medina. *State of the Art: Image Hashing*. <https://arxiv.org/abs/2108.11794>. 2021. arXiv: 2108.11794.
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. Vienna, Austria, 2021, pp. 1–21. DOI: 10.48550/arXiv.2010.11929.
- [8] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA, 2016, pp. 770–778.
- [9] K. Hosseini, D. C. S. Wilson, K. Beelen, and K. McDonough. “MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale”. In: *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities*. Seattle, WA, USA, 2022, pp. 8–19. DOI: 10.1145/3557919.3565812.
- [10] D. E. Knuth. *The Art of Computer Programming. Vol. 3, Sorting and Searching (2nd Ed.)* 2nd ed. Reading, MA, USA: Addison-Wesley, 1997.
- [11] Kopinor. *Bokhylla-avtalen (fra 2024)*. <https://www.kopinor.no/avtaletekster/bokhylla-avtalen-fra-2024>. 2023.

- [12] Y. A. Malkov and D. A. Yashunin. “Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.4 (2020), pp. 824–836. DOI: 10.1109/tpami.2018.2889473.
- [13] Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov. “Approximate Nearest Neighbor Algorithm Based on Navigable Small World Graphs”. In: *Information Systems* 45.null (2014), pp. 61–68. DOI: 10.1016/j.is.2013.10.006.
- [14] A. Mandal, S. Little, and S. Leavy. “Multimodal bias: Assessing gender bias in computer vision models with NLP techniques”. In: *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI ’23)*. Paris, France, 2023, pp. 416–424.
- [15] D. Mayo, J. Cummings, X. Lin, D. Gutfreund, B. Katz, and A. Barbu. “How hard are computer vision datasets? calibrating dataset difficulty to viewing time”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans, LA, USA, 2023, pp. 11008–11036.
- [16] T. Mei, Y. Rui, S. Li, and Q. Tian. “Multimedia Search Reranking: A Literature Survey”. In: *ACM Computing Surveys* 46.3 (2014), 38:1–38:38. DOI: 10.1145/2536798.
- [17] C. Meinecke, E. Guéville, D. J. Wrisley, and S. Jänicke. “Is Medieval Distant Viewing Possible? : Extending and Enriching Annotation of Legacy Image Collections Using Visual Analytics”. In: *Digital Scholarship in the Humanities* 39.2 (2024), pp. 638–656. DOI: 10.1093/llc/fqae020.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.null (2011), pp. 2825–2830.
- [19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Online, 2021, pp. 8748–8763.
- [20] N. Ruth, M. Burghardt, and B. Liebl. “From Clusters to Graphs - Toward a Scalable Viewing of News Videos”. In: *Computational Humanities Research Conference 2023 (CHR2023)*. Paris, France, 2023, pp. 167–177.
- [21] T. Smits and M. Kestemont. “Towards Multimodal Computational Humanities. Using CLIP to Analyze Late-Nineteenth Century Magic Lantern Slides.” In: *Computational Humanities Research Conference 2021 (CHR2021)*. Online, 2021, pp. 149–158.
- [22] T. Smits and M. Wevers. “A Multimodal Turn in Digital Humanities. Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections”. In: *Digital Scholarship in the Humanities* 38.3 (2023), pp. 1267–1280. DOI: 10.1093/llc/fqad008.
- [23] S. K. Snyderman, R. Sanderson, and T. Cramer. “The International Image Interoperability Framework (IIIF): A Community & Technology Approach for Web-Based Images”. In: *Archiving Conference*. Los Angeles, CA, USA., 2015, pp. 16–21.

- [24] K. Spärck Jones. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”. In: *Journal of Documentation* 28.1 (1972), pp. 11–21. doi: 10.1108/eb026526.
- [25] M. Wevers and T. Smits. “The Visual Digital Turn: Using Neural Networks to Study Historical Images”. In: *Digital Scholarship in the Humanities* 35.1 (2019), pp. 194–207. doi: 10.1093/lc/fqy085.
- [26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.
- [27] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. “Sigmoid Loss for Language Image Pre-Training”. In: *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*. Paris, France, 2023, pp. 11975–11986.
- [28] W. Zhou, H. Li, and Q. Tian. *Recent Advance in Content-based Image Retrieval: A Literature Survey*. <https://arxiv.org/abs/1706.06064>. 2017. arXiv: 1706.06064.

Bildesøk (beta)

Dette er en eksperimentell tjeneste fra DH-LAB som lar deg søke etter bilder i Nasjonalbibliotekets samling. Merk at denne tjenesten er under arbeid og kan endre seg. For øyeblikket støtter den bare baker fra før 1900-tallet.

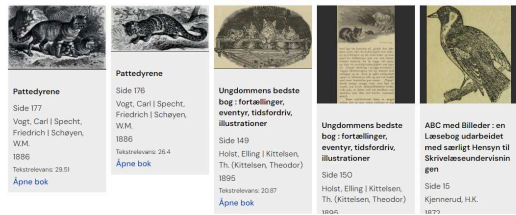
Hvordan fungerer søket?

Metode for søk: Skriv inn teksten

Skriv inn ord du vil bruke for å søke etter bilder:

Søkeinnstillinger

Bilder funnet for søket: kat

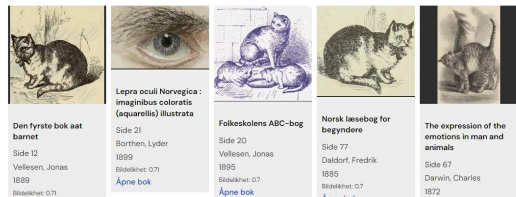


(a)

Søkeinnstillinger



Lignende bilder



(c)



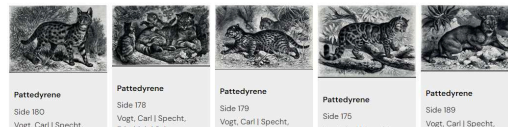
Pattedyrene

Side 177

Vogt, Carl | Specht, Friedrich | Schayen, WM, 1886

Åpne bok

Lignende bilder



(b)



The expression of the emotions in man and animals

Side 67

Darwin, Charles, 1872

Åpne bok







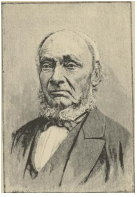

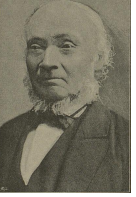



















Lignende bilder



(d)

Figure 1: Screenshots of the image search application: context-based search for "kat" (old Norwegian for cat) (a) and image-based query with a user-uploaded cat image (c). (b) and (d) show the results when selecting an image in (a) and (c), respectively. The app also has a collapsible sidebar (not shown) that we used for selecting SigLIP embedding vectors.

Table 2
 Example of search results using the different models

Query image	Model	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5
	SigLIP					
	CLIP					
	ViT					
	SigLIP					
	CLIP					
	ViT					

Continued on next page

















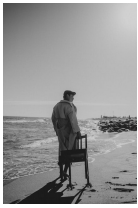






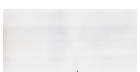








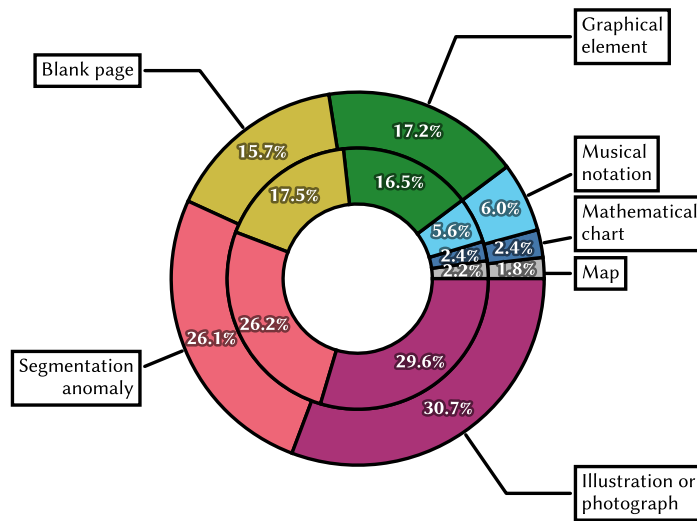
Query image	Model	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5
	SigLIP					
	CLIP					
	ViT					
	SigLIP					
	CLIP					
	ViT					

Table 3
Exact image retrieval accuracy

Accuracy Model	Top 1	Top 5	Top 10	Top 50
CLIP	492 (72 %)	596 (87 %)	613 (90 %)	638 (93 %)
SigLIP	529 (77 %)	633 (93 %)	645 (94 %)	665 (97 %)
ViT	529 (77 %)	582 (85 %)	597 (87 %)	612 (89 %)

Dataset Label	Train	Full*
Map	44	7692
Mathematical chart	48	10184
Musical notation	113	25398
Graphical element	330	72858
Blank page	349	66336
Segmentation anomaly	524	110254
Illustration or photograph	592	129867
In total	2000	422589

(a)



(b)

Figure 2: The class distribution for the manually labelled training set and estimated class distribution for the full dataset. (a) shows absolute counts, and (b) shows label distributions for the training set (inner) and estimated distributions for the full dataset (outer).

Table 4

Confusion matrix for the classification based on the outer cross-validation loop validation sets; it shows the number of elements with label a (columns) classified as label b (rows).

True class \ Predicted class	Segmentation anomaly	Blank page	Graphical element	Illustration or photograph	Musical notation	Map	Mathematical chart
Segmentation anomaly	496	5	28	8	2	1	2
Blank page	11	339	8	1	0	0	0
Graphical element	14	2	278	15	1	0	2
Illustration or photograph	1	3	16	558	1	2	5
Musical notation	1	0	0	0	109	0	0
Map	1	0	0	2	0	41	0
Mathematical chart	0	0	0	8	0	0	39

A perfect classifier will only have nonzero entries on the diagonal.