Visual Narratives: Large-scale Hierarchical Classification of Art-historical Images

Matthias Springstein[1,2] Stefanie Schneider[3] Javad Rahnama[4] Julian Stalter[3]
Maximilian Kristen[3] Eric Müller-Budack[1,2] Ralph Ewerth[1,2]
[1] TIB – Leibniz Information Centre for Science and Technology, Germany
[2] L3S Research Center, Leibniz University Hannover, Germany
[3] Ludwig Maximilian University of Munich, Germany [4] Reply GmbH, Germany

https://github.com/TIBHannover/iconclass-classification

Supplementary Material

This supplementary material further specifies the data sets used in the experiments (Appendix A), the network architecture and parameters (Appendix B), the performance of the pre-training with a smaller batch size (Appendix C), the performance of the hierarchical classification approaches (Appendix D), and details on the conducted qualitative analysis (Appendix E).

A. Data Sets

This section supplements the two data sets used in the experiments as discussed in Section 4 of the paper: (i) *ICAI (Iconclass AI Test Set)* [6]. The majority of the 362,561 annotations have a depth of five, six, or seven levels, with 109,701, 93,195, and 78,244 images, respectively. In particular, concepts with high granularity are rare; from level 10 onwards, only 1394 image examples are included. (ii) *ICARUS (Iconographic Classification and Representation Understanding)*. Among the 19 publicly available collections extracted by web-scraping, nine are from Germany (Artemis, Bildindex der Kunst & Architektur, Corpus Vitrearum, Heartfield Online, Hessen Kassel Heritage, Incunabulum Catalogue of the Bavarian State Library, Museen Thüringen, Städel Museum, Virtuelles Kupferstichkabinett),[1] two from Austria (Belvedere, REALonline),[2] one from Switzerland (Vitrosearch),[3] one from Poland (PAUart – Polish Academy of Arts and Sciences)[4], three from the Netherlands (Medieval Illuminated Manuscripts, Rijksmuseum, RKD – Netherlands Institute for Art History),[5] one from the United Kingdom (Broadside Ballads Online),[6] and two from the United States (Emblematica Online, National Gallery of Art).[7] Again, most of the 1,328,417 annotations have a depth of five, six, or seven levels, with 480,708, 410,854 and 301,731 images, respectively. However, the number of annotations with a depth of more than 10 levels is proportionally larger; here it is 9543 images.

Annotations from both *ICAI* [6] and *ICARUS* have been unified: (i) We remove ‘keys’ because many of them are redundant. This is partly because they are declared in lists and apply to a fixed range of *Iconclass* concepts, often just repeating aspects of the corresponding ‘keyless’ concept rather than differentiating it further (e.g., as in the *Iconclass* concept $C = \texttt{11D(+12)}$ with description $T_C =$ “Christ (+ Christ)”). (ii) We transfer *Iconclass* concepts with non-standardized ‘bracketed text’ to the next higher level of granularity in the taxonomy; automatic standardization is usually not readily feasible. As ‘non-standardized’ we consider named entities that are not regularly included in the *Iconclass* taxonomy, but can be added by the respective institutions. We refer to the *Iconclass* notation scheme in Fig-

[1] https://artemis.uni-muenchen.de/, https://www.bildindex.de/, https://corpusvitrearum.de/, http://heartfield.adk.de/, https://datenbank.museum-kassel.de/, https://inkunabeln.digitale-sammlungen.de/, http://www.museen.thueringen.de/, https://sammlung.staedelmuseum.de/, http://www.virtuelles-kupferstichkabinett.de/ (all last accessed on 2023-11-08).

[2] https://sammlung.belvedere.at/, https://realonline.imareal.sbg.ac.at/ (all last accessed on 2023-11-08).

[3] https://vitrosearch.ch/ (last accessed on 2023-11-08).

[4] http://pauart.pl/app (last accessed on 2023-11-08).

[5] http://manuscripts.kb.nl/, https://www.rijksmuseum.nl/, https://rkd.nl/ (all last accessed on 2023-11-08).

[6] http://ballads.bodleian.ox.ac.uk/ (last accessed on 2023-11-08).

[7] http://emblematica.library.illinois.edu/, https://www.nga.gov/ (all last accessed on 2023-11-08).

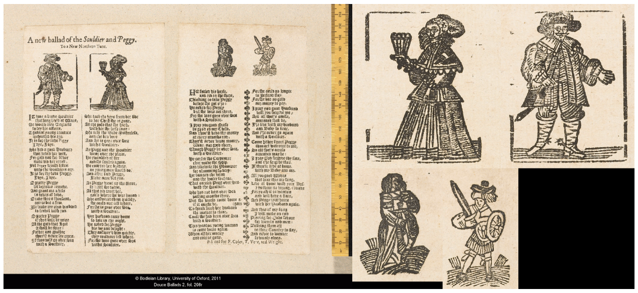
© Bodleian Library, University of Oxford, 2011
Douce Ballads 2, fol. 208r

Figure 1. Exemplary result of the two-stage pre-processing pipeline to normalize data sets.

ure 2a to illustrate the respective auxiliary components.

Many of the obtained images are scans containing extraneous noise or supplemental information, e.g., artist's signatures or linear color control charts; moreover, they may feature multiple artworks. To normalize the data sets, we introduce a two-stage pre-processing pipeline: (i) The relevant image content is first detected using a DeepLabv3 image segmentation model [1] with a ResNet-101 backbone [3]. For training, we annotated a small data set consisting of 101 images. (ii) The identified segments are then incorporated into a single rectangular segment by employing a 2D packaging algorithm [4]. This is necessary because *Iconclass* concepts were assigned for the entire image and cannot be attributed to individual segments. An exemplary result of this procedure is illustrated in Figure 1.

To exclude very similar images, we perform a near duplicate check on our proposed *ICARUS* data set. We use a ResNet-50 model [3] trained on the ImageNet-1K data set [7] to extract features for each image and compute the root mean square distance of the embeddings between all image pairs. After that, image pairs with a value below 0.5 were detected as duplicates; we kept one of the images and merged the annotations. Nevertheless, there are very similar images, as can be seen in Figure 7a of the paper, either because different artists reproduced the same subject or because preliminary drawings, sketches, and copies of the respective image exist.

B. Implementation Details

In addition to the parameters given in Section 5.1, to stabilize the training process, we use a linear ramp-up of the learning rate starting at 500 iterations and reduce it to zero at 40,000 iterations. Furthermore, we use a weight decay of 0.1 and prune the norm of all gradients to 0.5. For the CAT classification model (Section 3.4.5), the transformer decoder uses feature size $d_{model} = 768$ and a feed forward dimension of $d_{ff} = 2048$; the same configuration is employed by the vision transformer. Unless otherwise specified, we optimize our models on four Graphics Processing Units (GPUs) with 24GB RAM each.

Table 1. Results of contrastive pre-training with image-text pairs on different text generation strategies on the *ICARUS* test set using the CAT classifier with a batch size of 32. The results show the mean Average Precision (mAP) for all concepts that have at least one image in the test set. The best-performing strategy per batch size is denoted in bold.

Strategy	# of Training Images per *Iconclass* Concept			
	> 0	> 10	> 100	> 1000
KW	0.0476	0.0573	0.1155	0.2796
BLIP	0.0460	0.0554	0.1106	0.2700
GPT	**0.0560**	**0.0673**	**0.1314**	**0.3026**
LAION-400M	0.0426	0.0514	0.1049	0.2670

During the training process, we use the following image augmentation techniques: (i) images are padded to a square shape; (ii) random horizontal flip is applied; (iii) RandAugment [2] is applied with 2 operations and a size of 9; (iv) a 224 × 224 random section of each image is extracted that contains at least 70 % of the image. For testing, we padded the images and resized them to a dimension of 224 × 224.

C. Contrastive Pre-training with Image-Text Pairs with Smaller Batch Size

In addition to the batch size of 256 used in the other experiments, we also wanted to evaluate how the methods behave with a smaller batch size. To do this, we repeat the experiment from Section 5.3 with a batch size of 32. The results are shown in Table 1. While the descriptions generated by GPT perform best for a batch size of 32, the BLIP approach benefits from a larger batch size of 256 and provides the best overall results. A possible reason could be that the textual descriptions extracted with BLIP have less variance compared to GPT. As a result, the variance of negative prompts, which are important for training *CLIP* as mentioned in several works (e.g., [5]), may be too low using a small batch size.

D. Iconographic Concept Classification on Different Levels of Granularity

In these experiments, we compare our classification approaches presented in Section 3.4 with respect to the level of granularity in the *Iconclass* hierarchy. To do this, we average the mAP at each level on the *ICARUS* test data set, as shown in Figure 2.

As expected, performance drops considerably with higher levels of granularity, as the number of *Iconclass* concepts increases, while the number of training images per concept decreases. Interestingly, performance improves at levels eight and nine, especially for Flat-H and CAT. There are two reasons for this: firstly, the number of

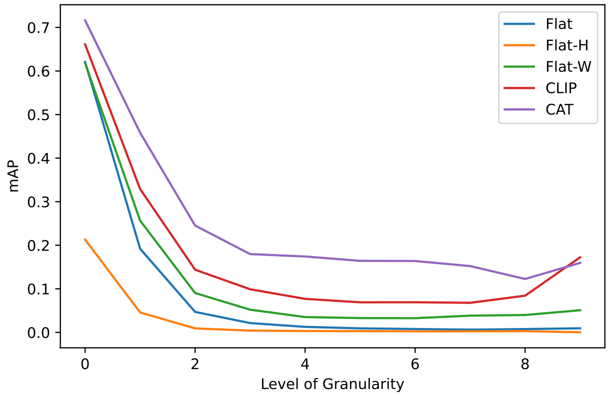

Figure 2. Results of the individual classification approaches on the *ICARUS* test set. The results show the mAP for all concepts at each level of granularity in the taxonomy.

Iconclass concepts in these stages decreases again, and secondly, both models optimize only those concepts that have a valid parent during training, thus reducing class imbalance. The higher performance of `Flat-H` compared to `CAT` might indicate that `CAT` can be enhanced with more iterations ($p = 30$), as the classifier might rarely reach the highest level of granularity.

E. Qualitative Evaluation

To qualitatively assess the `CAT` model's ability to make reasonable predictions, we recruited a total of four subjects. These included three art historians and one computer scientist proficient in relevant art-historical concepts. For the evaluation, we identified a set of 24 *Iconclass* concepts and filtered the top-10 classification results. Specifically, we selected four concepts each from six divisions within *Iconclass*, with the selected concepts within each division varying in their level of granularity. In selecting the concepts, care was taken to include as wide a range of art-historically significant narratives as possible, incorporating not only biblical themes, but also extending to narratives originating from, e.g., Greek mythology. For each concept, participants were asked to vote on whether they thought it was 'relevant' or 'irrelevant' to the concept in question, and to provide a brief comment to explain their assessment.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017. 2

[2] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Workshops co-located with the Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 14-19, 2020*, pages 3008–3017. IEEE, 2020. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition, CVPR 2016*, pages 770–778, New York, 2016. IEEE. 2

[4] Jukka Jylänki. A thousand ways to pack the bin. A practical approach to two-dimensional rectangle bin packing. 2010. 2

[5] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. arXiv:2212.00794, 2022. 2

[6] Etienne Posthumus. Iconclass AI test set. https://iconclass.org/testset/. Last accessed on 2023-11-08. 1

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2