Towards Multimodal Computational Humanities. Using CLIP to Analyze Late-Nineteenth Century Magic Lantern Slides

Thomas Smits, Mike Kestemont

University of Antwerp, Prinsstraat 13, 2000, Antwerpen, Belgium

Abstract

The introduction of the CLIP model signaled a breakthrough in multimodal deep learning. This paper examines whether CLIP can be fruitfully applied to a (binary) classification task in the Humanities. We focus on a historical collection of late-nineteenth century magic lantern slides from the Lucerna database. Based on the available metadata, we evaluate CLIP's performance on classifying slide images into 'exterior' and 'interior' categories. We compare the performance of several textual prompts for CLIP to two conventional mono-modal models (textual and visual) which we train and evaluate on the same stratified set of 5,244 magic lantern slides and their captions. We find that the textual and multimodal models achieve a respectable performance (∼0.80 accuracy) but are still outperformed by a vision model that was fine-tuned to the task (∼0.89). We flag three methodological issues that might arise from the application of CLIP in the (computational) humanities. First, the lack of (need for) labelled data makes it hard to inspect and/or interpret the performance of the model. Second, CLIP's zero-shot capability only allows for classification tasks to be simulated, which makes it doubtful if standard metrics can be used to compare its performance to text and/or image models. Third, the lack of effective prompt engineering techniques makes the performance of CLIP (highly) unstable.

Keywords

CLIP, classification, prompt engineering, multimodality, visual culture, magic latern slides,

1. Introduction

Following the development of deep learning models that are trained on expressions of a single sensory modality, mostly hearing (text) and seeing (images), researchers have recently focused on multimodal applications: models that process and relate information from multiple modalities [10]. While there are many different multimodal configurations, Baltrušaitis et al. (2019) note that text to image description (and, conversely, image to text), where the model is trained on image and text combinations, has emerged as the primary task of the subfield [2].

In January 2021, the introduction of the CLIP (Contrastive Language-Image Pre-training) signaled a breakthrough in the field of multimodal machine learning [12]. Trained on dataset of 400M image/text pairs collected from the internet, CLIP, given an image, must predict which out of a set of 32,768 randomly sampled text snippets it was paired with in the dataset. Radford et al. (2021) suggest that CLIP approaches this task by identifying visual concepts

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

thomas.smits@uantwerpen (T. Smits); mike.kestemont@uantwerpen.be (M. Kestemont)
0000-0001-8579-824X (T. Smits); 0000-0003-3590-693X (M. Kestemont)

in the images and associating them with textual descriptions [12]. As a result, the model can be applied to a wide variety of broad zero-shot 'text to image' and 'image to text' tasks.

While computer vision models have frequently been reported to outperform humans, they are optimized for performance on the specific task and data of the benchmark. As a result, their performance cannot be compared to the highly-contextual vision of humans [15]. Radford et al. (2021) report that CLIP matches the performance of computer vision models on thirty existing computer vision benchmarks, such as ImageNet, without being trained on the data of these benchmarks. CLIP thus shows a high performance 'in the wild' on tasks and datasets for which it was not optimized via training [12].

Building on recent discussions about the 'visual digital turn' [17], audio-visual Digital Humanities [1] and the connection between multimodality theory and digital humanities research [5, 6, 14, 18], this paper examines the application of a multimodal model to a (binary) classification task in the humanities. We focus on a historical collection of 40K magic lantern slides from the late-nineteenth century. The set includes digital reproductions of the slides (as a flat image), the title/captions (text), as well as meta-data (year of publication, mode of production). Recently recognized as being a highly multimodal medial form [8, 16, 19, 7], this collection of lantern slides provides an opportunity to evaluate the possible benefits of multimodal models for the (computational) humanities.

Based on the available metadata for the slides, we evaluate CLIP's performance on recognizing images of exterior/interior scenes. Seemingly purely visual in nature, multimodality theory would argue that text, such as captions, play a crucial role in producing these categories [3]. We compare the performance of CLIP to mono-modal text and image models, which we train and evaluate on a stratified set of 5,244 labelled magic lantern slides (and their captions) of exterior and interior locations. While the image model achieves the highest accuracy (∼0.898), we find that the best performing textual prompt for CLIP (interior/exterior) is competitive with the textual models (∼0.807 CLIP/∼0.806 BERT).

We flag three methodological issues that might arise from a possible widespread application of CLIP in the (computational) humanities. First, the lack of (need for) labelled data makes it hard to inspect and/or interpret the performance of the model. Second, even if labelled data is available, CLIP's zero-shot capability only allows for classification tasks to be simulated. As a result, it is doubtful whether accuracy and other standard metrics can be used to meaningfully compare CLIP to text and/or image models. Finally, the lack of methods to find the right, let alone the optimal, textual prompt(s) makes the performance of CLIP (highly) unstable. As a result, 'prompt engineering' [12] should be a major concern for future research that applies CLIP in the (computational) humanities.

This paper is part of the larger History of Implicit Bias project at the University of Antwerp, which applies machine learning to identify patterns of (implicit) bias in several nineteenth century digital collections. Multimodal machine learning could provide a breakthrough for this kind of research, which seeks to analyze large-scale and complex patterns of meaning in (historical) data. Models like CLIP could not only offer researchers the opportunity to study categories, such as 'the family,' that are highly multimodal in nature, but also, in conjunction with mono-modal techniques, fleece out the distribution of different modalities in meaning-making. This exploratory paper tests the robustness of CLIP to provide a sound basis for such research in the future.

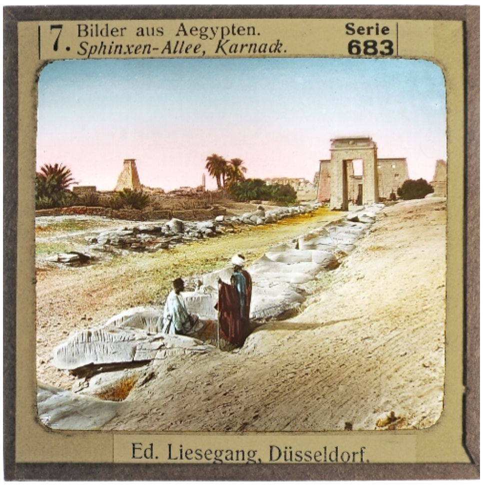


Figure 1: Example of exterior category
'Sphinxen-Allee, Karnack'
Slide 7 of *Bilder aus Ägypten* (year unknown).

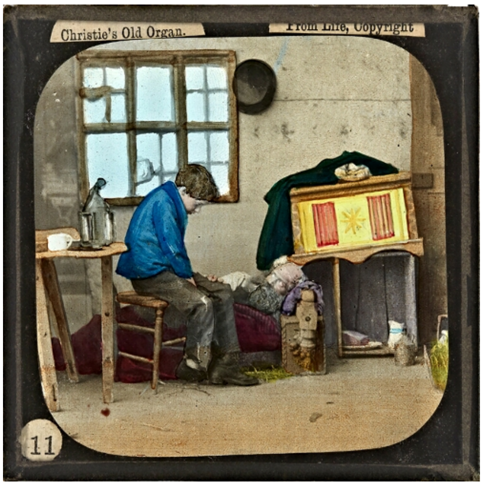


Figure 2: Example of interior category
'Christie tells Treffy only a month'
Slide 11 of *Christie's old organ* (1875).

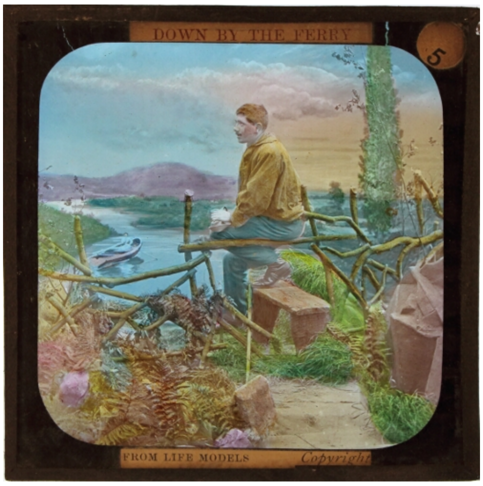


Figure 3: Example of a fictional exterior location
'Poor Robin cannot fly'
Slide 5 of *Down by the ferry* (1903).

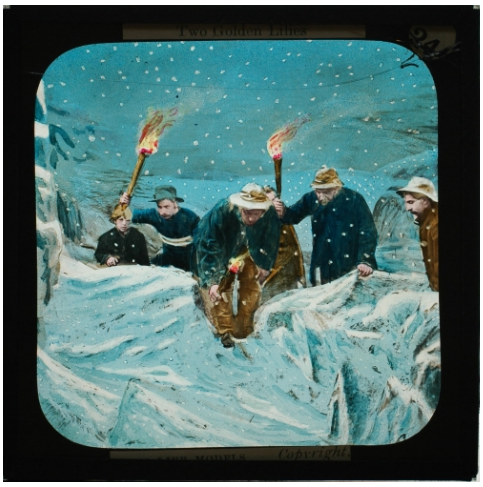


Figure 4: Example of a fictional exterior location
'The girl's footprints...'
Slide 24 of *The two golden lilies* (1893)

2. Material and methods

The study of the magic lantern has been stimulated by the increasing digital accessibility of lantern slides. The Lucerna Magic Lantern Web Resource was the first digital repository of digitized lantern slides. At the time of writing, it contained 42,019 digital slides, up from 38,000 in 2019, most of them uploaded and annotated by Lucerna's founder Richard Crangle [7]. We collected the digitized slides, their captions and several other metadata fields. The resulting dataset contains the URL, filename, title, year of publication, format, people connected to the slide, type of image, dimensions, materials, production process, person shown, image content tags, image location and collection for 42,019 slides (Dataset to be released with camera-ready paper).

To compare the performance of CLIP to mono-modal models on the exterior/interior classification task we used the ‘type of image’ field to produce a stratified .60/.20/.20 train, validation and test set of exterior and interior images with captions. As Table 1 shows, Lucerna’s slides were manually labelled for several types describing the physical setting captured on the slide. We combined the types ‘photograph of exterior location’ and ‘photograph of life models in exterior location’ to collect slides showing exterior locations (Fig. 1) and the ‘photograph of interior location’ and ‘photograph of life models in interior location’ types to collect slides of interior locations (Fig. 2). Initially, we also included the ‘photograph of life models in studio set’ in the collection of interior slides. However, as Fig. 3 and Fig. 4 show, this category often contains fictional ‘outdoor’ scenes [1] This demonstrates that seemingly binary categories, such as outdoor/indoor, often prove to be far-less rigid in actual practice. To enable comparison to a purely textual model, we only included slides with captions, discarding those without captions or with frequently recurring or generic ones, such as ‘Intro(duction)’ or ‘Title’. To create a balanced set, we included all the remaining slides of the interior category (2,622) and an equally-sized random sample of slides from the exterior category (5,244 total).

We compared the zero-shot performance of CLIP for several (apparently) binary prompts (Table 2) to a visual and a textual model (Table 3). The main advance of CLIP is that it does not need labeled training data to achieve competitive performance on a wide variety of classification tasks. However, this zero-shot capability results in the fact that we can only simulate a classification task. First, textual prompts have to be picked that are (apparent) mutually exclusive terms, phrases, or sentences. However, this does not exclude the possibility that both prompts are (un)likely textual descriptions of the same image. In contrast to models that are trained for a binary classification task, we do not ask CLIP a single question (Is this A or B?) but rather normalize the answers to two questions (Is this A?/Is this B?). Following earlier work, to calculate the accuracy of CLIP on a classification task, we use the softmax function to normalize the output of the model for the two prompts into a single probability distribution. While most deep learning models use softmax to normalize the output into a probability score, we ague that its application is conceptually different in the case of CLIP.

To compare CLIP’s zero-shot capabilities to mono-modal models we used relatively simple transfer learning methods. For the vision model, we applied the fast.ai framework to train a ResNet 18, a relatively simple convolutional neural network, pretrained on the ImageNet dataset. Instead of manually selecting hyperparameters, for example by determining the learning rate, we resorted to fast.ai’s default finetune method and its default parameters (for four epochs). For the text-only model, we first used a run-of-the-mill text classification approach [13], implemented in the established scikit-learn framework [11]. We represented the documents in train and test under a bag-of-words model. All features were normalized via the standard TF-IDF procedure (fitted on the training data only) to boost the weight of document-specific features. We report results for a word unigram model and a character trigram model. We applied a single-layer, linear classifier that is optimized via gradient descent to minimize a log loss objective. We have not optimized the hyperparameter settings and resort to default settings with an unpruned vocabulary (4,290 word unigrams; 6,358 character trigrams). The captions are primarily in English, but there some rare instances of other Western European languages (Dutch or German) which were not explicitly removed to increase the realism of the task.

[1] Copyright of Figures 1-4. Reproduced by permission via Lucerna Magic Lantern Web Resource. Figure 1: Private collection. Digital image © 2016 Anke Napp. Figure 2: The Hive. Digital image © 2018 Worcestershire County Council. Figure 3: Philip and Rosemary Banham Collection. Digital image © 2016 Philip and Rosemary Banham. Figure 4: Private collection. Digital image © 2006 Ludwig Vogl-Bienek.

Table 1
Absolute frequency distribution for the 'Type of image' field in Lucerna's full, original metadata. Only categories marked by an asterisk (*) were included.

Type of image	**Number of slides**
*photograph of exterior location	17,064
drawing / painting / print	11,789
photograph of life models in studio set	4,705
photograph	2,367
*photograph of interior location	2,075
*photograph of life models in exterior location	1,473
unknown	1,318
text	523
*photograph of life models in interior location	194
photograph of life models	174
NA	133
photograph of studio set	78
drawing / painting / print of exterior location	59
other	40
unknown of exterior location	20
physical object	6
text of exterior location	1
Total	42,019

We supplemented this, potentially naive, classifier with a generic, pretrained BERT for sequence classification. We started from the uncased, multilingual model from the Transformers library, which we finetuned as a binary exterior/interior classifier on the training set (we monitored on the development set via early stopping) and evaluated on the test set. The motivation for this was twofold. First, because we could bootstrap from a pretrained model, we expected the model to be able to model more subtle semantic aspects of the textual descriptions that aren't obvious from the lexical surface level (e.g. synonyms). Second, we started from the multilingual model that is available for this architecture: because our data is not exclusively monolingual, which could have given the BERT classifier a modest edge. A drawback of this neural approach is that model criticism through feature inspection is less straightforward.

3. Results

Table 3 shows that all models achieve a respectable accuracy, but that the vision-only model outperforms both CLIP and the textual models (almost a ∼50% error reduction). While it is a major advantage that CLIP does not require the labor-intensive and time-consuming process of producing labelled data and the training and fitting of models, the model is not competitive for this specific classification task. It depends on the questions of the humanities researcher whether a (possible) loss in accuracy is problematic. Researchers will have to make a decision whether possible improvements in accuracy warrant the investment needed to produce labelled data. Next to this pragmatic consideration, we argue that multimodal models also come with a new set of pitfalls. By comparing the performance of the text, vision and multimodal model, we flag three issues.

Table 2
Accuracy of prompts on exterior/interior categories

Prompts	accuracy on exterior	accuracy on interior	accuracy on all
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

Starting with the performance of CLIP, Table 2 shows that different prompts lead to different accuracy scores. From ∼0.96 for 'indoors' in the 'outdoors/indoors' prompt, to worse then guessing: ∼0.49 for 'outdoor' in the 'outdoor/indoor' prompt. Similar to the 'prompt engineering' discussions surrounding GPT-3 [4], Radford et al. (2021) note that determining the right prompt(s) can significantly improve the performance of CLIP [12]. The difference in accuracy between 'outdoor' and 'outdoors' (Table 2) is a good example of this.

In relation to prompt engineering, Radford et al. (2021) note that images are rarely paired with a single word [12]. As a result, they suggest that prompts that include contextual information achieve higher accuracy on several benchmarks. For example, 'a photograph of a German Sheppard, a type of dog' performs better then 'German Sheppard'. For our classification task, which seeks to distinguish between two high-level visual concepts, which are themselves already contextual, it is unclear what kind of information could improve the prompts. For example, the difference in performance between 'exterior/interior' and 'A photograph of an exterior/interior location' is limited (Table 2).

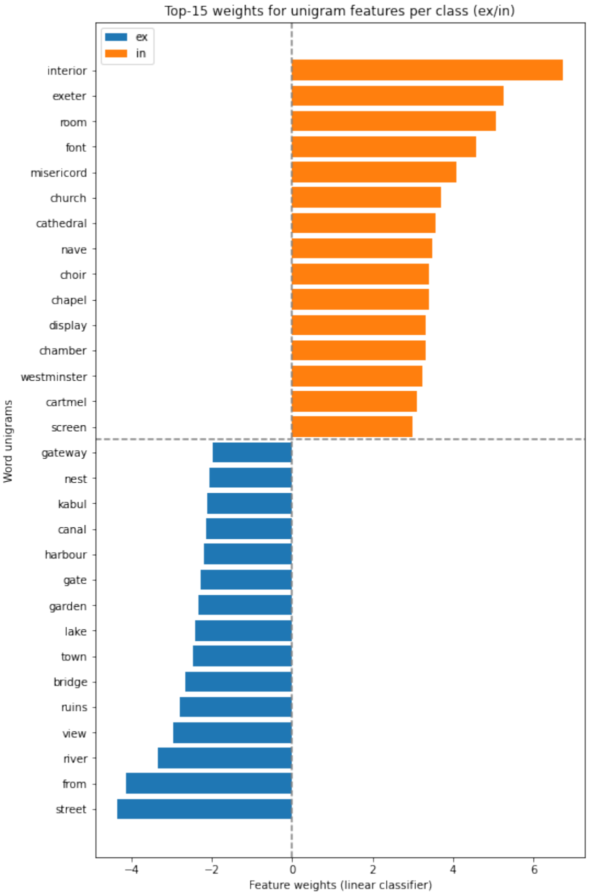


Figure 5: Top-scoring 15 weights for either class (ex/in) from the linear model for the token unigrams.

The limited increase in accuracy of adding 'a photograph of' to the prompts might be partly a result of the 'temporal bias' [15] of CLIP. The model was trained on 400M combinations of high-definition photographs and texts extracted from the internet. Although all the slides in our set are photographs, they look very different then the present-day images made by high-definition camera's. The fact that a large

Table 3
Accuracy of the textual, visual and multimodal models on the test set

Model	Description	Accuracy
Textual	Word unigrams	0.798
Textual	Character trigrams	0.777
Textual	BERT	0.806
Visual	ResNet 18, ImageNet weights	**0.898**
Multimodal	CLIP ('exterior/interior')	0.807

number of them are colored in (Fig. 1) might be the most striking visual difference. CLIP might not recognize (all) of our images as photographs, making it less beneficial to add this information to the prompts.

Looking at Table 2 we hypothesized that combining high performing words or snippets from different prompts might lead to better results. However, this is not the the case. While ‘exterior’ achieves high accuracy in the ‘exterior/interior’ prompt, its performance drops when combined with ‘indoors,’ which achieved high accuracy in the ‘outdoors/indoors’ combination and experiences an even more dramatic drop in accuracy when combined with ‘exterior’ (Table 2). This process can be explained by the fact that we normalize the output of the model for two prompts into a single probability distribution.

Regarding the textual models, a number of observations can be made. First, they score on par with the multimodal model, which is striking because the latter was trained nor finetuned on this specific dataset and task. Second, the visual model outperforms the textual models, suggesting that the textual modality is less relevant for this classification task. Interestingly, the word unigram model outperforms that based on character trigrams: this is an atypical result for a common text classification task and suggest that most of the useful cues in the title data is actually realised at the simple lexical level of atomic tokens. The visualization (Fig. 5) of the word unigram model’s highest weights for either class supports this hypothesis. Apart from the telltale feature ‘interior’, the indoor vocabulary is dominated by lexis related to the interior of church buildings (‘misericord’, ‘nave’, ‘choir’, etc.) – Exeter cathedral, in particular, might be over-represented in the data. The outdoor vocabulary, on the other hand, clearly points to more panoramic, landscape-related or aquatic (e.g. ‘bridge’, ‘lake’, ‘canal’, ‘harbour’) features or urban scenery (e.g. ‘street’, ‘town’, ‘gate’). The fixed expression ‘view from’ is also recognized by the model as a powerful lexical predictor of the exterior category. The fact that clear lexical clues are doing all the hard discriminatory work is also the suggested by the unimpressive performance of BERT: given its pretrained nature, in spite of the limited size of the training data set, we expected BERT to be able to harness at least some its pre-existing linguistic knowledge, but that hardly seems to be the case. Concerning prompt engineering, we hypothesized that highest weights for the two classes might result in relevant prompts for CLIP. However, as Table 2 shows, the combination street/interior does not lead to particularly good results.

Next to looking at the accuracy metric, we can use the top errors of CLIP and the visual model to compare them (Fig 6a). Clearly, the models have difficulties with different kinds of slides. The errors of the vision model seem the result from a lack of sky. The top error of CLIP is a result of mislabeling. While its caption (’*in* a Javanese home’) suggest the interior category, the image shows a family *outside* their house. CLIP wrongly attributed the other images to the exterior category, while they show details *inside* Exeter cathedral.

4. Discussion

Multimodal models hold the promise to lead to a 'practical revolution' in computational humanities research [9]. Instead of spending time (and money) on labelling datasets and training and fitting models, the zero-shot capabilities of CLIP could leave researchers free to apply deep learning techniques to more and different kinds of research questions and focus on the interpretation of results rather then the methods themselves. However, while CLIP has shown to be competitive on a large number of benchmarks, this paper demonstrates that this is not necessarily a given for all classification tasks. Relatively simple and easy to apply mono-modal models might significantly outperform CLIP for specific tasks. The fact that any textual prompt will yield *a* result, when not properly thresholded, might lead humanities scholars to expect too much. Future research should develop standardized practices to asses if results obtained with CLIP are reliable and meaningful. The fact that classification tasks can only be tackled indirectly, as we show in this exploratory paper, could pose a significant hurdle for future work. Traditional metrics, such as accuracy, might not be suitable to compare the performance of CLIP to other models. In line with this, the performance and reliability of CLIP could be significantly improved by better and more stable prompt engineering.

Acknowledgments

We would like to thank Ruben Ros for his help with the data collection and Melvin Wevers for helping setting up CLIP.

References

[1] T. Arnold, S. Scagliola, L. Tilton, and J. V. Gorp. "Introduction: Special Issue on AudioVisual Data in DH". In: *Digital Humanities Quarterly* 015.1 (2021).

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443. DOI: 10.1109/tpami.2018.2798607.

[3] J. A. Bateman. *Text and Image: A Critical Introduction to the Visual/Verbal Divide.* London; New York: Routledge, 2014.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models Are Few-Shot Learners". In: *arXiv:2005.14165 [cs]* (2020). arXiv: 2005.14165 [cs].

[5] T. Hiippala. "Distant Viewing and Multimodality Theory: Prospects and Challenges". In: *Multimodality & Society* (2021), p. 26349795211007094. DOI: 10.1177/26349795211007094.

[6] T. Hiippala and J. A. Bateman. “Semiotically-Grounded Distant Viewing of Diagrams: Insights from Two Multimodal Corpora”. In: *arXiv:2103.04692 [cs]* (2021). arXiv: 2103.04692 [cs].

[7] J. Kember. “The Magic Lantern: Open Medium”. In: *Early Popular Visual Culture* 17.1 (2019), pp. 1–8. DOI: 10.1080/17460654.2019.1640605.

[8] F. Kessler and S. Lenk. “Projecting Faith: French and Belgian Catholics and the Magic Lantern Before the First World War”. In: *Material Religion* 16.1 (2020), pp. 61–83. DOI: 10.1080/17432200.2019.1696560.

[9] B. Nicholson. “The Digital Turn”. In: *Media History* 19.1 (2013), pp. 59–73.

[10] L. Parcalabescu, N. Trost, and A. Frank. “What Is Multimodality?” In: *arXiv:2103.06304 [cs]* (2021). arXiv: 2103.06304 [cs].

[11] 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 E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[12] 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: *arXiv:2103.00020 [cs]* (2021). arXiv: 2103.00020 [cs].

[13] F. Sebastiani. “Machine learning in automated text categorization”. In: *ACM Comput. Surv.* 34.1 (2002), pp. 1–47. DOI: 10.1145/505282.505283. URL: https://doi.org/10.1145/505282.505283.

[14] T. Smits and R. Ros. “Quantifying Iconicity in 940K Online Circulations of 26 Iconic Photographs”. In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020).* Ed. by F. Karsdorp, B. McGillivray, A. Nerghes, and M. Wevers. Vol. 2723. Amsterdam: Ceur-ws, 2020, pp. 375–384.

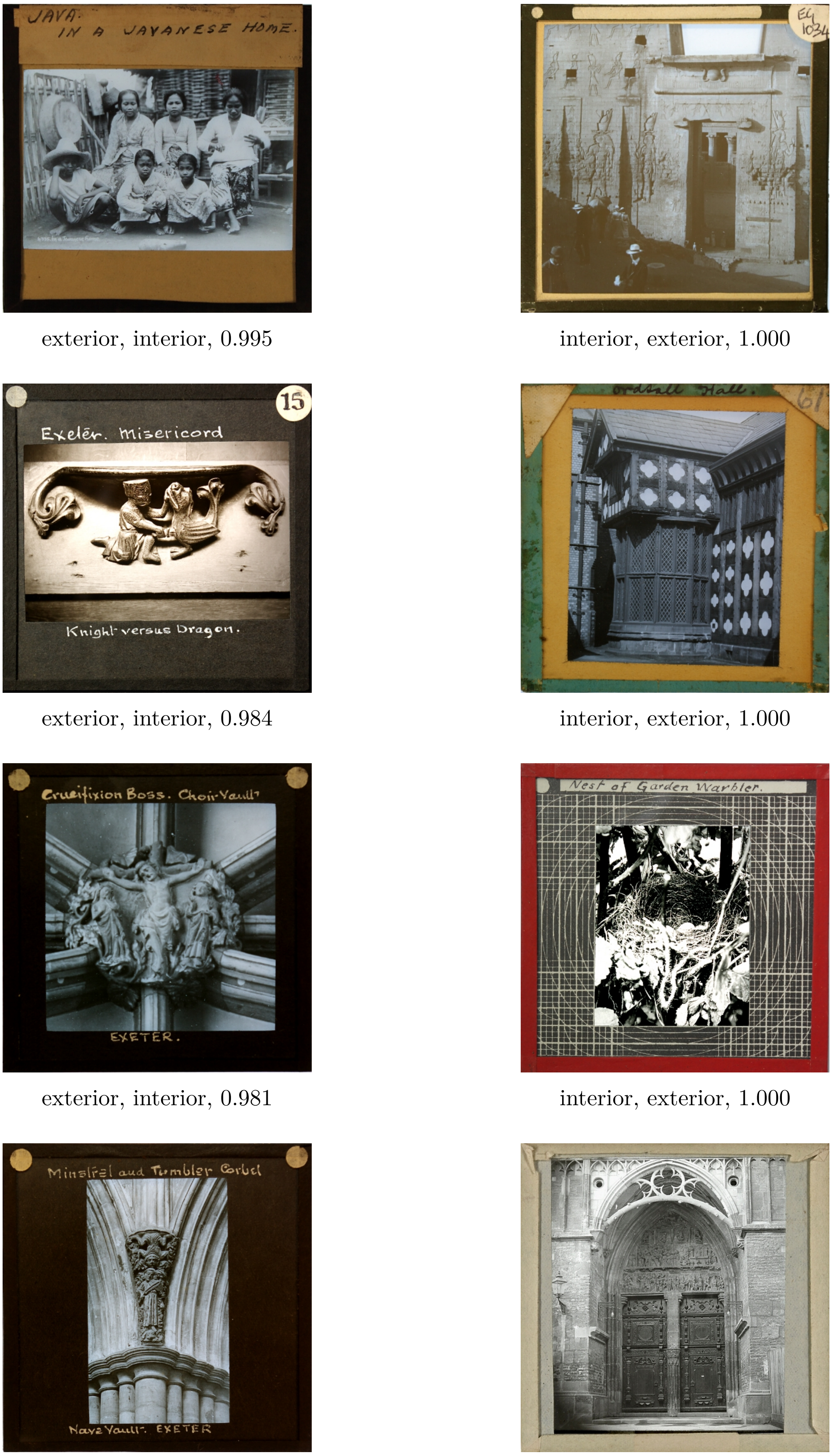
[15] T. Smits and M. Wevers. “The Agency of Computer Vision Models as Optical Instruments.” In: *Visual Communication* Online First (2021). DOI: 10.1177/1470357221992097.

[16] K. Vanhoutte and N. Wynants. “On the Passage of a Man of the Theatre through a Rather Brief Moment in Time: Henri Robin, Performing Astronomy in Nineteenth Century Paris”. In: *Early Popular Visual Culture* 15.2 (2017), pp. 152–174. DOI: 10.1080/17460654.2017.1318520.

[17] M. Wevers and T. Smits. “The Visual Digital Turn. Using Neural Networks to Study Historical Images”. In: *Digital Scholarship in the Humanities* 35.1 (2020), pp. 194–207. DOI: 10.1093/llc/fqy085.

[18] M. Wevers, T. Smits, and L. Impett. “Modeling the Genealogy of Imagetexts: Studying Images and Texts in Conjunction Using Computational Methods”. In.

[19] D. Yotova. “Presenting “The Other Half”: Jacob Riis’s Reform Photography and Magic Lantern Spectacles as the Beginning of Documentary Film”. In: *Visual Communication Quarterly* 26.2 (2019), pp. 91–105. DOI: 10.1080/15551393.2019.1598265.



exterior, interior, 0.995 | interior, exterior, 1.000

exterior, interior, 0.984 | interior, exterior, 1.000

exterior, interior, 0.981 | interior, exterior, 1.000

exterior, interior, 0.980 | interior, exterior, 1.000

Figure 6: Top 4 errors (prediction, actual, probability) for CLIP (col 1) and the visual model (col 2). Copyright Col 1 (top to bottom) reproduced by permission via Lucerna Magic Lantern Web Resource: Digital image © 2017 Manchester Museum/Digital image © 2014 Royal Albert Memorial Museum and Art Gallery, Exeter/Digital image © 2015 Royal Albert Memorial Museum and Art Gallery, Exeter/Digital image © 2015 Royal Albert Memorial Museum and Art Gallery, Exeter. Copyright Col 2 (top to bottom) reproduced by permission via Lucerna Magic Lantern Web Resource: Digital image © 2019 Manchester Museum/Digital image © 2015 Royal Albert Memorial Museum and Art Gallery, Exeter/Digital image © 2014 Royal Albert Memorial Museum and Art Gallery, Exeter/Digital image © 2018 Worcestershire County Council.