Viability of Zero-shot Classification and Search of Historical Photos

Erika Maksimova[1], Mari-Anna Meimer[1], Mari Piirsalu[1] and Priit Järv[1,*]

[1]*Institute of Software Science, Tallinn University of Technology, Estonia*

Abstract

Multimodal neural networks are models that learn concepts in multiple modalities. The models can perform tasks like zero-shot classification: associating images with textual labels without specific training. This promises both easier and more flexible use of digital photo archives, e.g. annotating and searching. We investigate whether existing multimodal models can perform these tasks, when the data differs from the typical computer vision training sets, on historical photos from a cultural context outside the English speaking world.

Keywords

zero-shot learning, digital heritage, multimodal models

1. Introduction

Cultural heritage archives may contain millions of photos. For efficient searching, the images need descriptions, like categories and captions. Traditionally, these are provided by human annotators. Ajapaik[1] is a crowd-sourced digital photo archive. It contains historical photos mainly from and related to Estonia and neighboring countries. The users of the archive upload and annotate the photos. Multiple collections from museums and the national archive have also been added. The earliest photos are dated before 1875, but the majority are taken from 1918 until present day. At the time of writing, Ajapaik contains 1181273 photos.

Figure 1 shows a screen capture from the website with image categorizations. The scene category can be either “exterior” or “interior”. The viewpoint elevation category is “ground”, “raised” or “aerial”. In the beginning of 2024, the scene category was specified for 43% and the viewpoint for 37% of the images. For the last three years, the number of images has been growing faster than the number of annotated images, meaning that the crowd of volunteers cannot keep up with the growth of the archive. The existing categories are somewhat limited and arbitrary, but adding new categorizations would further increase the annotation workload of the volunteers.

Training convolutional neural networks (CNN) to recognize the categories has been the conventional approach to automated annotation of images. Such task specific models have asso-

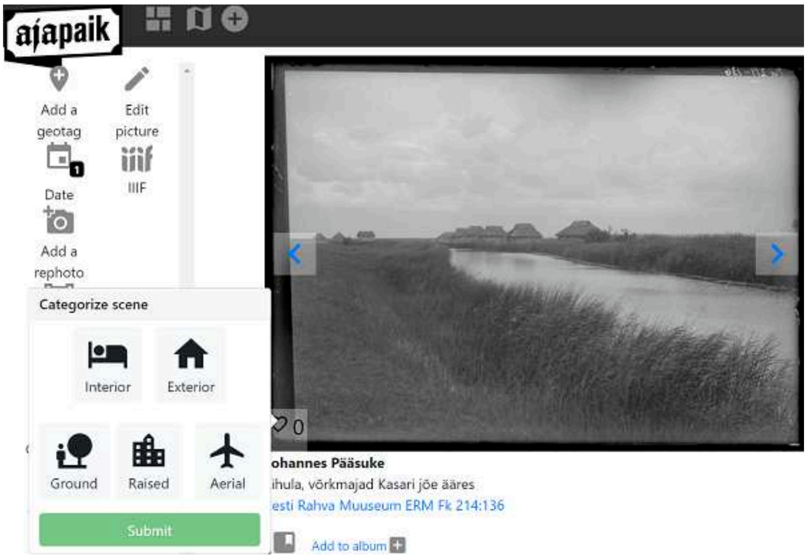
CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

erika.maksimova@gmail.com (E. Maksimova); marianna.meimer@taltech.ee (M. Meimer); mari.piirsalu@taltech.ee (M. Piirsalu); priit.jarv1@taltech.ee (P. Järv)

0000-0001-7725-543X (P. Järv)

[1] https://ajapaik.ee/



ohannes Pääsuke
ihula, võrkmajad Kasari jõe ääres
esti Rahva Muuseum ERM Fk 214:136

Figure 1: Ajapaik photo view with image categorizations. Photo CC BY 4.0, Johannes Pääsuke, ERM Fk 214:136, Eesti Rahva Muuseum, https://opendata.muis.ee/object/605721

ciated costs – the human labor involved in preparation of the training data and the operation and maintenance of the model.

Pre-trained multimodal neural networks come with the promise of removing these costs. As an example, CLIP [9] is a relatively lightweight model that can encode both natural language input and images into a shared multimodal vector representation. For example, the text "cow" and a picture of a cow would have a very similar representation. This allows the model to perform zero-shot classification: when presented with an example of an image and a label, the model can immediately predict whether the label is associated with the image, without needing to see any other examples. The implication is that CLIP and other similar multimodal models can replace supervised computer vision models, like CNNs, without requiring large training data sets.

There are limits to how well machine learning algorithms generalize to unseen data. Ultimately, a model can only learn some representation of its training data, so the usefulness of the model depends on how similar the distribution of data in the application is to the distribution of the training data. For example, a study by de Vries et al. measured a 15-20% difference in the classification accuracy of household items like soap between images from the United States and from Somalia and Burkina Faso [15]. They associate this with representational bias, as the majority of images in many computer vision datasets originate from Western countries. Historical appearance of locations, items, situations and common activities is similarly underrepresented. Therefore, existing successful use cases of multimodal models on modern photographs, or their evaluations on standard computer vision datasets like ImageNet [10] and CIFAR100 [5] are not reliable indicators of their usefulness for historical photos.

In this paper, we investigate if using off-the-shelf multimodal models is a viable method for classifying photos from Ajapaik. To understand what trade-offs or drawbacks this involves, we do a comparison with supervised computer vision models. The multimodal vector representations from the CLIP model can also be adapted for searching images based on their visual content. This would be a very useful functionality for image collections, so we include the evaluation of multimodal search in our experiment.

2. Related Work

The main inspiration for our work comes from the paper by Smits and Wevers [12]. They demonstrate the capabilities of the CLIP model [9] on collections of magic lantern slides and children's books illustrations, originating from the 19th century to approximately 1940. In the task of classifying indoor and outdoor images, CLIP was slightly less accurate than a convolutional neural network trained specifically for the task. Smits and Wevers identify several forms of bias that cause the model to make mistakes, like mis-identifying modern concepts in historic images, and applying sex-role stereotypes. Their finding is that the differences in visual representation do not impact the performance negatively. As an example, the concept of family is recognized from illustrations of both people and anthropomorphic animals.

The report on CLIP by Radford et al. [9] also evaluates the model's zero-shot performance against a fully supervised neural network model. CLIP with no task-specific training outperforms ResNet50 [3] that was separately trained for each specific task, in classifying several frequently used computer vision datasets, like ImageNet and CIFAR100. Radford et al. caution that their evaluation set may be co-aligned with the capabilities of the model, which means that the high performance is not guaranteed to carry over to applications. For example, in their paper CLIP underperforms in the specialized task of classifying satellite images. The authors observe that the natural language interface may be unsuited to specify more complex tasks.

We refer to the papers by Aske and Giardinetti [1] and Männistö et. al [8] for a wider overview of machine learning for visual archives and omit the discussion of methods that do not involve large multimodal models here. Few papers explicitly investigate multimodal models for historical images classification and retrieval. Barancová et al. explore the dating of historical photos, concluding that zero-shot classification is relatively inefficient [2]. They achieved better results by training a classifier on top of the multimodal model's image representation. Tschirschwitz et al. propose evaluation datasets and a framework for historical images classification and retrieval [14]. In their study, zero-shot CLIP achieves the highest performance, however they also report that additional qualitative evaluation did not confirm their quantitative results. Springstein et al. present methods of classification of art-historical images in a hierarchical schema of visual themes [13]. Their work does not include off-the-shelf multimodal models in the zero-shot setting.

Our paper uses the CLIP, SigLIP [17] and BLIP-2 [6] models that combine vision and language transformers and can encode images and text into a shared multimodal vector representation. There are many more multimodal models, with over 60 models cited in a recent survey [16]. This number is constantly growing. Most of these models are designed for "downstream" tasks like image and text generation, and do not necessarily provide documented interfaces for classification or for accessing shared multimodal representations.

The contribution of our paper, as compared to evaluations in [9], is that we use a dataset that is superficially similar (photos of people, everyday life, buildings) to mainstream computer vision training datasets but different in two aspects. The pictures are from different historical eras, and from outside the English speaking cultural sphere, which is significant because the representations of concepts in models are learned through language modelling. Our contribution is complementary to [12] and [14] as it is a similar investigation on a different dataset. We report results for both classification and search. We break down the search evaluation to differentiate between very general concepts, and entities and objects that are distinctly local to the cultural context of Ajapaik.

3. Methods

We present two experiments, covering two common use cases of a photo collection: classification and search. The source code of the experiments is available at https://github.com/priitj/chr2024/. Because the copyright of most of the photos used in the experiments is held privately, we cannot reproduce the photos in the paper and do not distribute the datasets.

Our classification experiment measures the capability of multimodal models to automatically label photos in a collection with a fixed set of categories. In the search experiment, a query text is given and a multimodal model is used to retrieve a set of matching images. We measure how well the models rank the images by the relevance of their visual content to the query text.

We evaluate three multimodal models. CLIP was used in previous research on historical images [12, 2, 14, 13], and is generally widely adopted and cited, so we include it as a reference multimodal model. SigLIP is very similar to CLIP, but uses a different training objective. In the evaluation done by it's authors, SigLIP outperforms different variations of CLIP in all classification tests [17]. Based on this, we included SigLIP in our study. For both of these models, we use HuggingFace Transformers[2] implementations.

BLIP-2 is another model with an architecture optimized for efficient training. It is reported to outperform CLIP in text-to-image retrieval [6]. For this model we used implementations from the Salesforce LAVIS[3] language-vision package.

The nomenclature of these models used in machine learning literature also includes the description of their vision transformer component. A ViT-B vision transformer is the "base" size with 12 transformer layers, while a ViT-L is the "large" transformer, with 24 layers and increases in other settings as well. The *patch size* describes how the input image is partitioned before feeding it to the transformer: a "patch 32" model uses 32x32 pixel rectangles. Therefore, the input fed to the transformer of a 16x16 patch model is actually four times larger, making it more computationally expensive. These variations have an impact on the performance of the models, so we include three different configurations of both CLIP and SigLIP in our experiments.

[2] https://huggingface.co/

[3] https://github.com/salesforce/LAVIS

Table 1
The number of images in the classification set

Category	Category Label	Photos
Scene	interior	3056
	exterior	8614
Viewpoint elevation	ground	10662
	raised	3057
	aerial	3053

3.1. Classification Experiment

In the classification experiment, we use a sample of 17042 photos from the Ajapaik collection. The images are annotated with the scene category, the viewpoint elevation category, or both. Table 1 gives the number of photos in each category. Additional photos were added to less frequent categories like "raised", such that each category has at least 3000 photos. This was done to obtain better performance with the supervised baseline models and to ensure enough test examples in those categories. For the remainder of the paper, we refer to this sample as the *classification set*.

We test scene category and viewpoint elevation category classification separately. With the supervised baselines, we use 5-fold cross-validation. In each round of cross-validation, the images are split 75:5:20 between train, validation and test parts. The 5% validation part is only used to automatically select the best model during training. All the measurements in reported in the paper are done on the 20% test parts, which cover the entire set of images in a given category over the 5 rounds of cross-validation.

With CLIP and SigLIP, classifications for all images in a given category are computed directly by the model from the input of an image and a set of prompts. We use the category labels "interior", "exterior", "ground", "raised" and "aerial" as initial prompts for their respective categories.

Both Radford et al. [9] and Smits and Wevers [12] observe that the selection of prompts to represent the classes has a noticeable effect on classification performance. With the Ajapaik classification set, this effect should also be expected. The single word "raised" does not describe the class very precisely and only becomes meaningful if we know that the context is viewpoint elevation. Accordingly, we expand our sets of prompts for classification with different natural language phrases describing the categories (Tables 6 and 8).

To the best of our knowledge, BLIP-2 does not include a classification model. We implement a simple nearest neighbors classifier on top of BLIP-2 vector representation. We compute image vectors I_i and text vectors T_j with BLIP-2 for an image i and a prompt j. The class of the image is the one that maximises the similarity of vectors:

$$sim(i, j) = \frac{I_i \cdot T_j}{\|I_i\| \|T_j\|} \tag{1}$$

For baselines, we use convolutional neural networks (CNNs) and transfer learning. All selected models are pre-trained on ImageNet. We train a shallow classifier on top of the pre-trained CNNs. Scene category classifiers and viewpoint elevation classifiers are trained separately.

For CNN architectures, we selected ResNet18 and ResNet50, as they were used as baselines in papers [12] and [9], respectively. Additionally, we selected DenseNet121 [4] to represent a deeper architecture, and MobileNetV2 [11] as a more modern, lightweight computer vision model.

We measure the classification performance using per-class F1-score. For a given class, true positives (TP) is the number of images that were correctly predicted by the model to be in this class. False positives (FP) is the number of images that were incorrectly predicted to be in the class. False negatives (FN) is the number of images that belong to the class, but the model predicted a different class label. The F1-score penalizes both false positives and false negatives:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

3.2. Search Experiment

For the search experiment, we downloaded the metadata of photos from the Ajapaik API.[4] We then randomly selected 11000 photos that had a textual description. Because of downloading and image file format errors, our final sample has 10846 images. We will refer to this sample as the *search set.*

We evaluate the search performance by letting the models rank the images by their relevance to search terms. To ensure that at least one relevant photo exists for each search term, we extracted the search terms from the textual descriptions of the images in the search set. We translated descriptions of images to English using the deep-translator[5] package. We then POS tagged and lemmatized the words in the descriptions, and detected named entities using the SpaCy[6] library.

We select the search terms with the assumption that the users would mostly search for objects (such as “boat”), events (“exhibition”), activities (“riding”) or named entities, like a place name or a person. These search terms would also be non-ambiguous enough so that we can decide whether a retrieved photo is relevant to the search term. We select English language nouns as examples of objects, events and phenomena. Verbs are examples of activities. In total, we used 8 different categories, with 10 terms in each category. The selected search terms are listed in Table 2.

Common and rare sets are included so that both easier and more difficult searches are represented. In common sets, we selected 10 terms that occurred most frequently in descriptions. All common search terms had occurred with at least 20 photos, and sometimes with hundreds. In rare sets, we selected 10 random terms from among those that had occurred once.

The random English words selection is used to diversify the search terms and to reduce any unintentional bias from the frequency based selection of other terms. The same random words are human translated to Estonian. This set is used to test whether the models give any useful results with non-English input.

[4] https://opendata.ajapaik.ee/

[5] https://github.com/nidhaloff/deep-translator

[6] https://spacy.io

Table 2
Selected search terms

	Common	Rare
Objects, events, phenomena	view, farm, building, group, anniversary, portrait, exhibition, child, school, competition	venue, wetland, stomach, tricolor, rumba, detective, lore, heath, population, score
Activities	commissioning, performing, leaving, speaking, taking, giving, working, making, sitting, standing	fertilizing, hoeing, windmilling, watering, mourning, seamstressing, illustrating, schoolmastering, stitching, binding
Named entities	Tallinn, the University of Tartu, Tartu, Viljandi, Narva, Tartu University, Harju, V. Kingissepa, Rakvere, Moscow	Jaan Kadakas, Jüri Randla, U.K. Kekkonen, Ralf Allikvee, A. Rosenberg, Setumaa, Karl Parts, Sara Teitelbaum, Margit Tooman, Valeri Kirss
	English	Estonian
Random words	winter, interior, milk, village, boy, driving, education, fence, lecturer, horse	talv, interjöör, piim, küla, poiss, sõitmine, haridus, aed, lektor, hobune

The terms selected by automatic criteria required some manual changes. For example, with named entities, we decided not to include names of countries, because "Estonia" would match with the majority of photos. An excluded search term was replaced with the next most frequent term for the common categories, and a new randomly selected term for the other categories.

We implement search by computing text-to-image similarity. For an image i and search term j, the similarity is computed using Equation 1 from the multimodal representation vectors I_i and T_j. The search results for a search term j are k most similar images, sorted by $sim(i, j)$. Our implementation uses the Voyager[7] approximate nearest neighbors index to find and sort the most similar images.

We evaluate the ability of models to rank relevant results above irrelevant ones. We selected mean average precision (MAP) as the measurement of the quality of search results [7, p. 155-161]. MAP is a robust measure that is not sensitive to the number of relevant documents in the search set. If there are not enough relevant photos, MAP does not penalize filling the remainder of the search results with irrelevant photos.

We report the measurements for the top-k results, considering that this is what the user will see in practical application. For a query q belonging to a set of queries Q, let $R_q(k)$ be the set of relevant results among the first k results. Let r_i be a search result at position i. Average precision for k results, or AP@k, is calculated

$$\text{AP@}k(q) = \frac{1}{|R_q(k)|} \sum_{r_i \in R_q(k)} \frac{|R_q(i)|}{i} \tag{3}$$

[7] https://github.com/spotify/voyager

Table 3
Classification performance of supervised baselines

Model	F1				
	Interior	Exterior	Ground	Raised	Aerial
ResNet18	0.87	**0.96**	0.92	0.63	0.90
ResNet50	**0.89**	**0.96**	**0.94**	0.70	**0.92**
DenseNet121	0.87	**0.96**	0.92	0.66	**0.92**
MobileNetV2	0.86	0.95	0.93	**0.71**	0.91

AP rewards ranking relevant results above irrelevant results. For example, if the first 4 among top 10 results are relevant, then AP@10 = 1.0. If the 4 relevant results are ranked below 6 irrelevant results, $\text{AP@10} = \frac{1}{4}\left(\frac{1}{7} + \frac{2}{8} + \frac{3}{9} + \frac{4}{10}\right) \approx 0.28$. If there are no relevant results, AP = 0. Mean average precision (MAP) is calculated over a set of queries:

$$\text{MAP@}k(Q) = \frac{1}{|Q|} \sum_{q \in Q} AP@k(q) \tag{4}$$

While the search set contains some positive labels to evaluate relevance of images to queries, there are no negative labels. For example, if a description of a photo includes the word “boy” we could count it as relevant towards a query “boy”, however there is no information about what the photo *does not* depict. Therefore, the relevance of each photo in search results to each search term was evaluated by human judges.

4. Results

We begin with the performance of the supervised baselines to provide a frame of reference to the results obtained with multimodal models. Table 3 lists the classification results with the CNN models trained specifically for the scene category and viewpoint elevation classification tasks. We observe that firstly, the “interior”/“exterior” classification is the easier one of the two tasks shown. Secondly, the “raised” category is ambiguous as a textual description, but based on Table 3 it is also ambiguous visually. The per-class F1-score for “raised” is 20-30 percentage points lower than other classes, indicating that the supervised models struggle generalizing this concept.

With multimodal models, we first report per-class classification performance when using class labels “as is”. For example, when classifying the viewpoint elevation category, the image is matched with the texts “ground”, “raised” and “aerial”.

Table 4 lists the per-class F1 scores for scene and viewpoint elevation categories. As with the supervised baselines, the scene category classification works better. Also similarly to the supervised baselines, the “raised” class is the most difficult. However, compared to Table 3, the multimodal models do much worse, with F1-scores ranging from 0.02 to 0.24.

Several outcomes were unexpected. Contrary to the evaluations by the authors of the models [9, 17], smaller versions of the models perform equal or better than bigger ones. Unlike in the paper by Zhai, et al. [17], SigLIP does not clearly outperform CLIP in classification, in fact

Table 4
Classification performance with class labels as prompts

Model	Transf. size	Patch size	Interior	Exterior	F1 Ground	Raised	Aerial
CLIP	ViT-B	32	**0.82**	**0.94**	**0.82**	0.06	0.62
	ViT-B	16	0.81	0.93	0.79	0.02	0.64
	ViT-L	14	0.78	0.90	0.79	0.11	0.73
SigLIP	ViT-B	16	0.76	0.88	0.77	**0.24**	**0.80**
	ViT-L	16	0.77	0.89	0.74	0.17	0.74
	SO-400m	14	0.79	0.90	0.76	0.05	0.73
BLIP-2	ViT-L	14	0.65	0.78	0.80	0.20	0.69

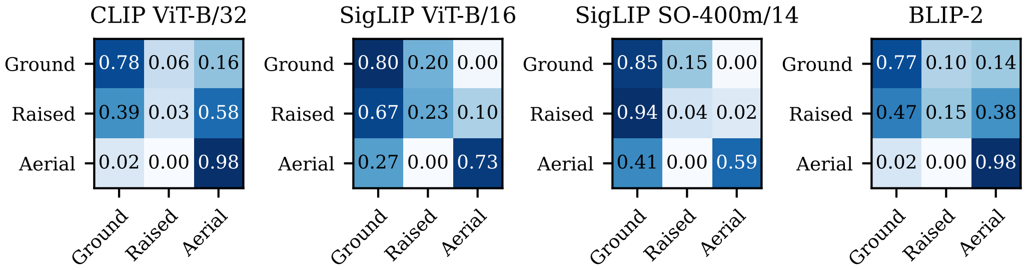


Figure 2: Confusion matrices for viewpoint elevation. Rows: true classes, columns: predicted classes.

the small CLIP ViT-B/32 model is the best in predicting three classes out of five. It is also surprising that the performance in the aerial category is low, because we would expect the aerial photographs to be visually distinct.

The reason behind the lower performance for the aerial category is revealed by the confusion matrices in Figure 2. The rows are the true classes and the columns are the classes that the model predicted. The CLIP ViT-B/32 and BLIP-2 models labeled 98% of aerial pictures correctly, only 2% of them were labeled as ground. Their low performance is caused by the false positives, as they heavily tend towards the aerial category and also label other photos as aerial.

In comparison, the SigLIP models tend heavily towards predicting the ground category. Due to having fewer false positives for aerial, the per-class score is higher. The trade-off is that the per-class score for ground is lowered. SigLIP ViT-B/16 has the most balanced performance in the viewpoint elevation category, thanks to being able to detect raised elevation photos better than the other models.

Using more descriptive prompts allowed multimodal models to reach higher performance, but the overall impact of prompt engineering was mixed. We provide the full results in Appendix A. The Tables 6–7 give the prompts and mean F1-scores for the scene category classification task. The Tables 8–9 are the prompts and results for the viewpoint elevation classification task.

A brief summary: with the prompts “indoor scene” and “outdoor scene”, SigLIP ViT-B/16

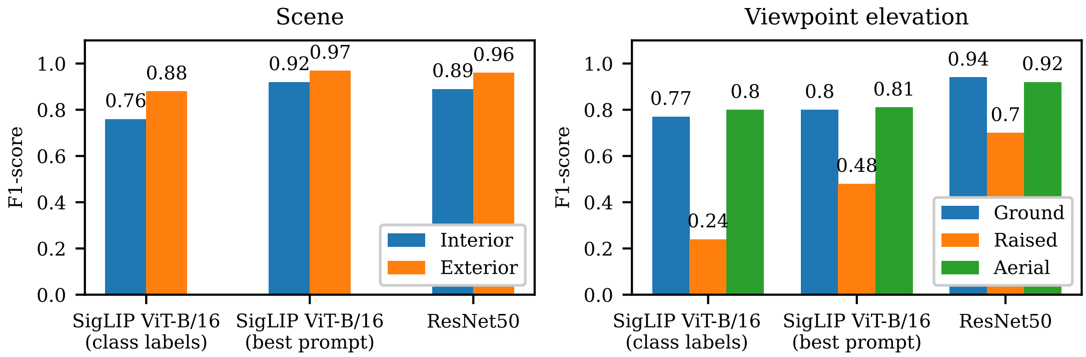


Figure 3: Comparison of multimodal and supervised baseline classifiers.

model achieves F1 = 0.92 for the interior and F1 = 0.97 for the exterior scenes. This is the only instance out of all combinations of models and prompts, where a multimodal model outperforms any of the supervised baselines. However, out of 42 tests using the more descriptive prompts in scene category classification, in 24 instances the performance dropped, compared to using class labels directly. The more descriptive prompts had a positive impact for the CLIP model in viewpoint elevation classification. In total, we did 84 tests with prompt engineering and in exactly half (42) the performance improved or remained the same. In the remaining 42 test instances the performance dropped.

Figure 3 compares the best performing multimodal models and supervised baselines. We selected SigLIP ViT-B/16 as the representative of multimodal models, as it achieved multiple highest per-class and mean F1-scores, including the one result that outperformed the baselines. We selected ResNet50 as the representative of CNNs. The best results of multimodal models are competitive with supervised baselines in scene classification and below the baselines by a large margin in viewpoint elevation classification.

The search experiments are summarized in Table 5. Two weak categories are clearly visible – the rare named entities and the Estonian language search terms. The models perform the best when searching for objects, then activities and are overall weakest when searching for named entities.

Like in the classification experiment, model performance relative to each other differs from evaluations in previous literature. Surprisingly, the best performing model is clearly SigLIP SO-400m/14, which is optimized for classification [17]. BLIP-2, which we included due to prior strong results in text-to-image retrieval, did not perform as well.

The MAP scores in Table 5 do not tell us directly whether there were many correct results and the ranking within top-10 did not matter, or if the models were able to precisely rank relevant photos above irrelevant ones. We analyze the ranking ability in the left graph of Figure 4. The AP@10 results are plotted against the number of relevant photos in top-10. With SigLIP SO-400m/14, the best performing model, we still see that the best AP@10 score drops below 1.0 when there were fewer than 4 relevant results. In other words, in searches with 1-3 matches, there were always irrelevant photos ranked above relevant ones.

Table 5
Search results by search term categories, mean average precision (MAP@10)

Model	Transf. size	Patch size	Objects Com.	Objects Rare	Activities Com.	Activities Rare	Named Ent. Com.	Named Ent. Rare	Random Eng.	Random Est.
CLIP	ViT-B	32	0.83	0.45	0.79	0.47	0.39	0.05	0.85	0.19
	ViT-B	16	0.85	0.40	0.68	0.60	0.47	0.01	0.89	0.04
	ViT-L	14	0.83	0.43	0.69	0.47	0.54	**0.06**	0.90	0.23
SigLIP	ViT-B	16	0.92	0.53	0.79	0.63	0.39	0.01	0.90	0.40
	ViT-L	16	0.94	0.55	0.79	**0.75**	0.50	0.00	0.92	0.39
	SO-400m	14	**0.95**	**0.59**	**0.84**	0.73	**0.74**	0.02	**0.96**	**0.57**
BLIP-2	ViT-L	14	0.84	0.47	0.71	0.47	0.19	0.05	0.85	0.12

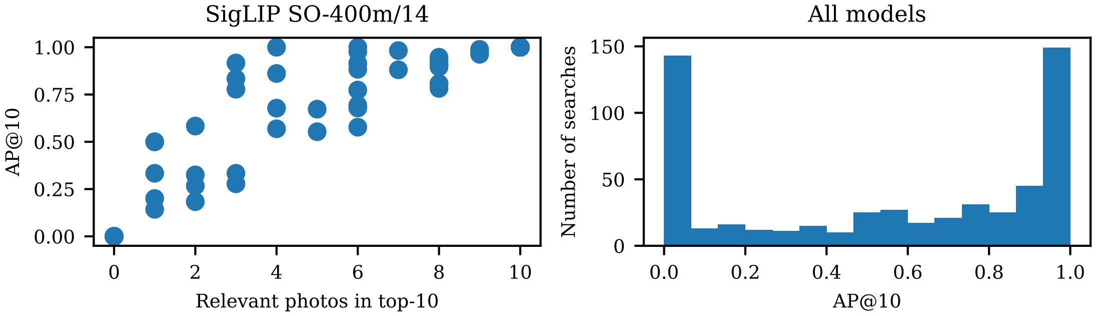


Figure 4: Left: ranking performance, AP@10 by the number of relevant results in top-10. Right: overall distribution of AP@10.

The second question is, whether the models can find all search terms, or are there gaps that the aggregate MAP scores do not show. We present the distribution of all AP@10 scores in Figure 4, right graph. Over all models, the most likely result of a search is either a complete success or a complete failure, with intermediate results much less likely.

Clearly, many of the low AP@10 results come from the named entity or Estonian language searches. When we look at the other five categories of regular English language words, a more interesting result emerges. For each model, there are 3-10 terms where AP@10 was under 0.3. At the same time, for each term at least one model achieved $AP@10 > 0.5$, except the word "score" where the best $AP@10 = 0.38$. Therefore, each model has gaps in the knowledge, but these gaps lie in different places for different models, including different sizes of CLIP and SigLIP.

5. Discussion

Prior evaluations on mainstream computer vision datasets do not generalize well to the case of Ajapaik. Previously established rankings of which models perform better in classification and search ([17] and [6], respectively) were not reproduced in our experiments. This implies that if the users want to ensure good performance, they need to test their own particular use case, which would involve human evaluation or annotation of data.

The multimodal models did well in the scene category classification task and less so in viewpoint elevation classification. Prompt engineering closed the gap to supervised baselines, but the multimodal models responded to prompts unpredictably. Using more descriptive prompts, like “elevated view” instead of “raised”, was equally likely to increase or to decrease the performance. This is important, because there was no prior indication of what model and prompt set would be best. We only know that SigLIP ViT-B/16 performed well in scene category thanks to having the annotated classification set.

The evidence from prompt engineering tests shows that the difficulties encountered in the classification have more to do with having to specify the classification task precisely through the natural language interface, as was in fact anticipated by Radford et al. when they discuss applications of CLIP [9]. The only practical way of mitigating this is to move away from the zero-shot setting. Training a classifier on top of multimodal representations, like done in [2], may require similar amounts of annotated data like with CNNs. However, Radford et al. show improved performance with multi-shot learning where much fewer examples are needed (up to 16 in the paper) [9].

The impact of the cultural sphere on the search results was clear. The Estonian language, localities and persons are not well represented in the models. When we remove this requirement of out of domain knowledge and look at 50 search terms of common English words, the models still have gaps, with 3-10 searches per model failing. Importantly, which terms fail differs by model. On the positive side, this unpredictability is not as pronounced as with the classification. The top search results are clearly populated with relevant photos for common objects, activities and random English worlds, independent of the model.

Our paper has multiple limitations. The relevance of a photo to a given search term in search results was validated by one person each. Methodologically it would be preferable if several persons validated one result. However, the photo–term pairs were distributed between the judges in a way that was, for practical purposes, random. Each set of photos returned by a model was therefore validated by multiple judges, which should dilute the effect of possible bias.

In the search experiment, we used the Voyager approximate nearest neighbors index. It is possible that Voyager had some impact on the search results, if it did not return the exact k nearest neighbors set each time. To validate this, the search experiment has to be repeated without an index, and we omitted this because of additional labor needed to validate the results.

There are many other multimodal models that we did not include in our experiment. Additional investigation is needed to determine, which of those, if any, provide an interface to shared multimodal vector representation of text and images. Finally, we used only one dataset, so our result may not generalize to other similar datasets.

6. Conclusions

We investigated the viability of zero-shot classification and search of historical photos in Ajapaik. We found that this application domain is different enough from the models’ training data that expectations based on previous evaluations on model performance do not hold.

Multimodal models can successfully search for common everyday concepts from the photos. However, the zero-shot usage on this archive is problematic. Firstly, models have unpredictable gaps in knowledge with common English words that appear depending on model size. Secondly, knowledge of Estonian language words and names is mostly missing. Thirdly, classification performance is below supervised baselines and cannot be easily improved with prompt engineering. Therefore, in the context of historical visual archives, the multimodal models do not deliver on the promise of removing dataset annotation related costs. Thorough evaluation is recommended to ensure viability in each use case.

Acknowledgments

We thank the anonymous reviewers for their comments that helped to improve the paper. We are grateful to Anna Grund for providing the annotated classification set and to the contributors of Ajapaik, who made writing this paper possible.

References

[1] K. Aske and M. Giardinetti. "(Mis)Matching Metadata: Improving Accessibility in Digital Visual Archives through the EyCon Project". In: *ACM Journal on Computing and Cultural Heritage* 16.4 (2023), 76:1–76:20. DOI: 10.1145/3594726.

[2] A. Barancová, M. Wevers, and N. van Noord. "Blind Dates: Examining the Expression of Temporality in Historical Photographs". In: *Proceedings of the Computational Humanities Research Conference*. Ed. by A. Sela, F. Jannidis, and I. Romanowska. Vol. 3558. CEUR Workshop Proceedings. Paris, France: CEUR-WS.org, 2023, pp. 490–499. URL: https://ceur-ws.org/Vol-3558/paper5790.pdf.

[3] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Las Vegas, NV, USA: IEEE Computer Society, 2016, pp. 770–778. DOI: 10.1109/cvpr.2016.90.

[4] F. N. Iandola, M. W. Moskewicz, S. Karayev, R. B. Girshick, T. Darrell, and K. Keutzer. *DenseNet: Implementing Efficient ConvNet Descriptor Pyramids*. arXiv:1404.1869 [cs.CV]. 2014. DOI: 10.48550/arXiv.1404.1869.

[5] A. Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. Technical Report, University of Toronto. 2009. URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[6] J. Li, D. Li, S. Savarese, and S. C. H. Hoi. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". In: *International Conference on Machine Learning, ICML*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. Proceedings of Machine Learning Research. Honolulu, Hawaii, USA: Pmlr, 2023, pp. 19730–19742. URL: https://proceedings.mlr.press/v202/li23q.html.

[7] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge, UK: Cambridge University Press, 2008. DOI: 10.1017/cbo9780511809071.

[8] A. Männistö, M. Seker, A. Iosifidis, and J. Raitoharju. *Automatic Image Content Extraction: Operationalizing Machine Learning in Humanistic Photographic Studies of Large Visual Archives.* arXiv:2204.02149 [cs.CV]. 2022. DOI: 10.48550/arXiv.2204.02149.

[9] 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, ICML.* Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. Virtual Event: Pmlr, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *Int. J. Comput. Vis.* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[11] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* Salt Lake City, UT, USA: Computer Vision Foundation / IEEE Computer Society, 2018, pp. 4510–4520. DOI: 10.1109/cvpr.2018.00474.

[12] 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: *Digit. Scholarsh. Humanit.* 38.3 (2023), pp. 1267–1280. DOI: 10.1093/llc/fqad008.

[13] M. Springstein, S. Schneider, J. Rahnama, J. Stalter, M. Kristen, E. Müller-Budack, and R. Ewerth. "Visual Narratives: Large-scale Hierarchical Classification of Art-historical Images". In: *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV.* Waikoloa, HI, USA: Ieee, 2024, pp. 7195–7205. DOI: 10.1109/wacv57701.2024.00705.

[14] D. Tschirschwitz, F. Klemstein, H. Schmidgen, and V. Rodehorst. "Drawing the Line: A Dual Evaluation Approach for Shaping Ground Truth in Image Retrieval Using Rich Visual Embeddings of Historical Images". In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing, HIPICDAR 2023.* San Jose, CA, USA: Acm, 2023, pp. 13–18. DOI: 10.1145/3604951.3605524.

[15] T. de Vries, I. Misra, C. Wang, and L. van der Maaten. "Does Object Recognition Work for Everyone?" In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops.* Long Beach, CA, USA: Computer Vision Foundation / IEEE, 2019, pp. 52–59. URL: http://openaccess.thecvf.com/content%5C%5FCVPRW%5C%5F2019/html/cv4gc/de%5C%5FVries%5C%5FDoes%5C%5FObject%5C%5FRecognition%5C%5FWork%5C%5Ffor%5C%5FEveryone%5C%5FCVPRW%5C%5F2019%5C%5Fpaper.html.

[16] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. *A Survey on Multimodal Large Language Models.* arXiv:2306.13549 [cs.CV]. 2023. DOI: 10.48550/arXiv.2306.13549.

[17] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. “Sigmoid Loss for Language Image Pre-Training”. In: *IEEE/CVF International Conference on Computer Vision, ICCV*. Paris, France: Ieee, 2023, pp. 11941–11952. DOI: 10.1109/iccv51070.2023.01100.

Table 6
Prompts for scene classification

Prompt set	Interior	Exterior
P_{11}	interior	exterior
P_{12}	interior scene	exterior scene
P_{13}	interior view	outdoors
P_{14}	indoor scene	outdoor scene
P_{15}	indoors	outdoors
P_{16}	inside a building	outside
P_{17}	inside a room	outside

Table 7
Scene category classification with different prompt sets

Model	Transf.	Patch	Mean F1-score						
	size	size	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}
CLIP	ViT-B	32	**0.88**	0.87	0.58	0.92	0.79	0.79	0.88
	ViT-B	16	0.87	0.88	0.37	0.76	0.71	0.73	0.86
	ViT-L	14	0.84	0.79	0.47	0.86	0.69	0.76	**0.92**
SigLIP	ViT-B	16	0.82	**0.92**	0.76	**0.95**	0.76	**0.81**	0.76
	ViT-L	16	0.83	0.65	0.69	0.94	0.69	0.80	0.74
	SO-400m	14	0.85	0.91	0.71	0.94	0.78	0.69	0.65
BLIP-2	ViT-L	14	0.71	0.80	**0.81**	0.88	**0.87**	0.80	0.83

Table 8
Prompts for viewpoint elevation classification

Prompt set	Ground	Raised	Aerial
P_{21}	ground	raised	aerial
P_{22}	ground view	raised view	aerial view
P_{23}	street level	view from building	view from airplane
P_{24}	ground view	elevated view	bird's eye view
P_{25}	ground	elevated view	aerial
P_{26}	ground level	elevated view	aerial
P_{27}	ground level	elevated view	aerial view

A. Prompt Engineering

The appendix contains the prompt sets and the corresponding results for scene classification (Tables 6–7) and viewpoint elevation classification (Tables 8–9).

Table 9
Viewpoint elevation classification with different prompt sets

Model	Transf. size	Patch size	Mean F1-score P_{21}	P_{22}	P_{23}	P_{24}	P_{25}	P_{26}	P_{27}
CLIP	ViT-B	32	0.50	0.40	0.51	0.50	**0.68**	**0.70**	**0.73**
	ViT-B	16	0.49	0.58	0.57	0.60	0.65	0.66	0.68
	ViT-L	14	0.54	0.61	**0.67**	0.55	0.67	**0.70**	0.70
SigLIP	ViT-B	16	**0.60**	**0.70**	0.54	0.49	0.30	0.29	0.34
	ViT-L	16	0.55	0.52	0.52	**0.69**	0.49	0.52	0.66
	SO-400m	14	0.51	0.49	0.59	0.61	0.39	0.39	0.55
BLIP-2	ViT-L	14	0.56	0.50	0.49	0.54	0.61	0.47	0.51