2023 RISP International Workshop on Nonlinear Circuits, Communications and Signal
Processing (RISP NCSP 2023), Feb. 28 - Mar. 3, 2023

Visual Archive Search Using Vision-language Object Detection Models

Ryuichi Tomiya, Tristan Hascoet, Ryoichi Takashima and Tetsuya Takiguchi

Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe, 657-8501, Japan
Phone/FAX:+81-78-803-6570
E-mail: 1945061t@gsuite.kobe-u.ac.jp, tristan@people.kobe-u.ac.jp,
rtakashima@port.kobe-u.ac.jp and takigu@kobe-u.ac.jp

Abstract

Recently, by adding natural language understanding to computer vision, object detection models can detect nuanced concepts from free-form text without specific training. A visual search is important to understanding accidents and the extent of the damage in natural disaster response and planning. In this paper, we discuss the use of a vision-language object detection model to carry out a visual search of an image archive of an earthquake disaster. We will demonstrate the effectiveness of a vision-language object detection model in the field of object detection by varying the level of complexity of the image-related text to detect objects such as “a backpack”, “a safety cone”, “a blue tarp”, “a person sitting”, “a person wearing a helmet”, and “a person riding a bike”. We will also compare the accuracy of vision-language object detection models and an open-vocabulary image classification model in visual searches and analyze the tendency.

1. Introduction

Many applications in science and industry require the extraction of specific information from a large collection of data. One such application is natural disaster response and planning. In such an application, a visual search is important to understanding accidents and the extent of damage.

In a typical visual search using computer vision, for example, to detect “a person riding a bike”, the images of “a person riding a bike” must be prepared and used to train an object detection model to detect an image containing “a person riding a bike”. It is not realistic to prepare a large amount of training data for each such query. On the other hand, by adding natural language understanding to computer vision, detection becomes possible without specific training.

In this paper, we demonstrate the effectiveness of vision-language object detection pre-trained models of GLIP [1] and MDETR [2] in a visual search using the digital archive of the Great Hanshin-Awaji Earthquake Disaster. We also use an open-vocabulary image classification pre-trained model of CLIP [3], and compare the performances of these classification and detection models in a visual archive search.

2. Related work

CLIP, pre-trained by large-scale image-text pairs, make open-vocabulary image classification possible. Inspired by CLIP, vision-language object detection models, such as GLIP and MDETR, have been proposed. They detect nuanced concepts from free-form text, and generalize them to unseen combinations of categories and attributes. Recently, image retrieval using CLIP has been proposed in the fields of fashion [4], commerce [5], and medicine [6]. We evaluate text-based visual archive searches using GLIP and MDETR.

3. Methods

A visual search is carried out using the Great Hanshin-Awaji Earthquake Disaster Materials Collection Digital Gallery owned by Kobe University. In order to investigate the vision-language object detection pre-trained models to determine their object search efficiency, in our experiments, the 24,303 images from the archive (without annotations) and the text query describing the object being searched for are input into the GLIP (GLIP-L) and MDETR (EfficientNet-B5) models, and the CLIP (ViT-B/32) model. The retrieved images are then arranged in order of decreasing output score to evaluate the visual archive search. In a visual archive search, filtering is necessary to pick images that contain the requested object and to discard images that do not contain the requested object. An illustration of a random human search pipeline is shown in Figure 1 (left). An illustration of a visual archive search using a vision-language object detection pre-trained model pipeline is shown in Figure 1 (right).

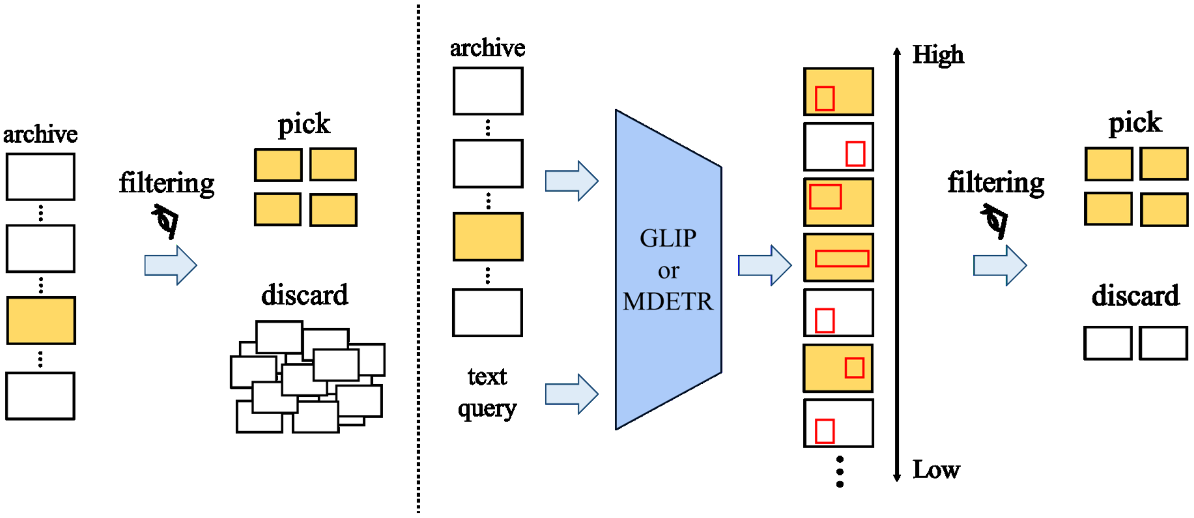


Figure 1: Random human search pipeline (left) and visual archive search using vision-language object detection model pipeline (right)

4. Results

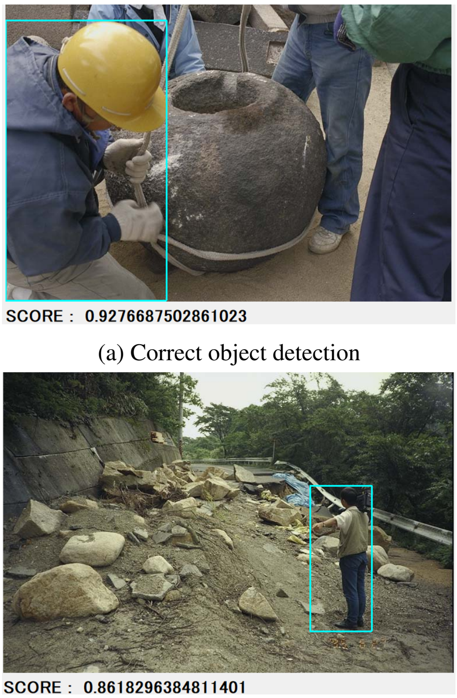
4.1 Experiment on the 24,303 images (without annotations)

In order to evaluate the visual archive search, we define "visual search efficiency gain" in our experiments as follows:

$$
\begin{aligned}
&visual\ search\ efficiency\ gain \\
&= \frac{Precision^{model}(TP=50)}{Precision^{human}(TP=50)} \qquad (1) \\
&= \frac{\frac{50}{TP+FP(model)}}{\frac{50}{TP+FP(human)}} = \frac{TP+FP(human)}{TP+FP(model)}
\end{aligned}
$$

where TP represents the number of true positives and $Precision^{model}(TP = 50)$ and $Precision^{human}(TP = 50)$ represent the precision score of GLIP, MDETR or CLIP, and by-human (without GLIP, MDETR or CLIP) when $TP = 50$, respectively. This efficiency gain means that if the gain is greater than 1, the vision-language object detection or classification is effective in comparison to a random human search.

Table 1 shows the visual search efficiency gain for GLIP, MDETR and CLIP for six queries. Using GLIP and MDETR, they are arranged in order of decreasing output score. Also, using CLIP, it calculates the cosine similarity of the image and the text query, and then they are arranged in order of decreasing similarity. Their images, which contain the detected object, are employed for the evaluation of the visual archive search. On the other hand, in the random human search, 600 images are randomly selected and a person detects the object in the images for each query. The number of the detected objects for "a backpack", "a safety cone", "a blue tarp", "a per-



(a) Correct object detection

(b) Incorrect object detection

Figure 2: Examples for "a person wearing a helmet" detected by GLIP (photo by Yoshimichi Ohgimoto (Kobe University Library Great Hanshin-Awaji Earthquake Disaster Materials Collection))

Table 1: Comparison of visual search efficiency gain

	GLIP	MDETR	CLIP
“a backpack”	**11.58**	2.31	3.34
“a safety cone”	**3.94**	1.09	2.44
“a blue tarp”	3.43	4.45	**7.11**
“a person sitting”	1.51	1.50	**5.77**
“a person wearing a helmet”	1.04	1.15	**2.87**
“a person riding a bike”	1.45	1.72	**9.70**

son sitting”, “a person wearing a helmet” and “a person riding a bike” is 32, 99, 63, 65, 85, and 34, respectively. Therefore, for example, $TP + FP(human)$ for “a backpack” is $(600/32) \times 50 = 938$.

As can be seen from the GLIP and MDETR gains in Table 1, vision-language object detection is effective for visual archive searches for all six queries. In particular, a simple text query was highly effective in searching for “a backpack”, “a safety cone”, and “a blue tarp”. But the efficiency gain for the query “a person ___ing” is not so high. A possible reason for this is that there are many cases where “a person” is detected, but “a person ___ing” is not detected. Figure 2 shows examples of the detection output. The top image is correctly detected and the bottom image is not correctly detected, where “a person” is correctly detected but “a person not wearing a helmet” with a high output score. In addition to the example of “a person not wearing a helmet” getting a high output score, there were many examples of high scores for confusing objects, such as “a person wearing a cap, a hat”.

In addition, a comparison of GLIP and MDETR show that GLIP tended to be more effective in searching for “a backpack” and “a safety cone”.

As can be seen in the CLIP gains in Table 1, open-vocabulary image classification is also effective for visual archive searches for all six queries. Efficiency gain for “a backpack” and, “a safety cone” is low, but the efficiency gain for the query “a person ___ing” is high. Incorrect images that have a high similarity, they tend to have characteristics different from those obtained from the detection models. Figure 3 is the 4th-ranked image out of 24,303 images when sorted by “a person sitting”. The image shows a chair, but not “a person”. Since CLIP calculates the similarity between the entire image and the text, it may be characterized as having a high similarity to “images of scenes or places where the object indicated by the text is likely to be in the image”.

Comparing the detection models with the classification model, we found that the classification model is superior in finding objects among a large number of images, in regard to “a person ___ing”.

4.2 Experiment on the 600 images (with annotations)

Next, a visual search was performed on 600 images that were labeled by a human. The purpose of this experiment was to assess the ability to find every object in the data set without any omissions.

Table 2 shows the Average Precision, a measure of whether all 600 images were efficiently sorted or not. The same tendency as in Table 1 is observed for “a backpack” and “a safety cone”, where GLIP is most effective. However, GLIP is also the most effective at finding “a blue tarp” and “a person riding a bike”, which differs from the tendency in Table 1. It is thought that the reason for this may be that GLIP is effective in finding all small objects without omission.

Table 2: Comparison of Average Precision [%]

	GLIP	MDETR	CLIP
“a backpack”	**19.94**	6.07	13.58
“a safety cone”	**84.78**	27.59	28.94
“a blue tarp”	**36.81**	28.24	22.57
“a person sitting”	18.52	22.39	**25.06**
“a person wearing a helmet”	9.25	6.26	**17.27**
“a person riding a bike”	**17.61**	11.10	16.15

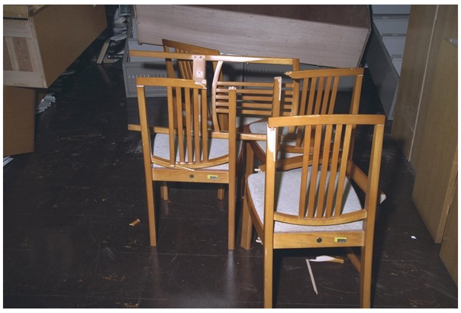


Figure 3: Example of an image with a high similarity to “a person sitting” (photo by Kobe University Library (Kobe University Library Great Hanshin-Awaji Earthquake Disaster Materials Collection))

We hypothesize that the main subject of an image impacts the accuracy because a classification model (CLIP) is trained to text associated with the main subject of an image.

Here, the main subject of an image was related to the percentage of the image occupied by the object. Figure 4 plots the total percentage of an image that is occupied by the requested object, showing that CLIP was less effective than GLIP for “a backpack”, “a safety cone”, and “a person riding a bike” compared to the other objects, which occupy a smaller percentage of the image than the other objects. On the contrary, “a person sitting” and “a person wearing a helmet”, for which CLIP was more effective than GLIP, occupy a larger percentage of the image than the other objects. However, although “a blue tarp” occupies a large percentage of the image, CLIP is not as effective as GLIP in locating it. It is clear that the accuracy of CLIP does not always increase when the percentage of the image occupied by the requested object is high.

The fact that a large percentage of an image is occupied by an object may not necessarily indicate that the object is the main subject of the image. In the random human search, “a backpack”, “a safety cone” or “a blue tarp” tended not to be the main subject of an image. Conversely, “a person __ing” tended to be the main subject of an image. This suggests that whether or not an object is the main subject of an image has a significant impact in the effectiveness of the search. However, since the main subject of an image cannot be quantitatively determined, it is difficult to test this hypothesis.

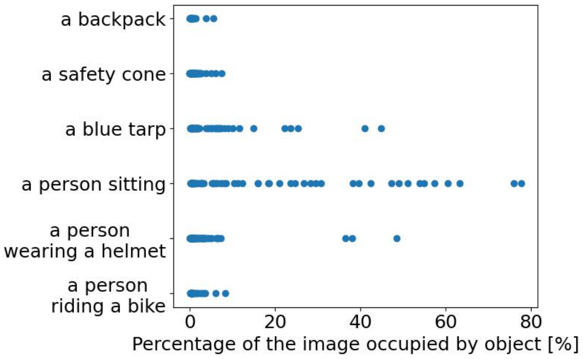


Figure 4: Distribution of the percentage of the image occupied by objects in the 600 images

5. Conclusions

We evaluated vision-language object detection models by determining their efficiency in visual archive searches using text (phrase) queries. We compare the performances of classification and detection models in a visual archive search. The relationship between the main subject of an image and the search accuracy is not clearly understood. In future research, we would like to fine-tune our method using the true positive information obtained from our evaluation.

Acknowledgment

We are grateful to Kobe University for allowing us to use Kobe University Library Great Hanshin-Awaji Earthquake Disaster Materials Collection in our research.

References

[1] Li, Liunian Harold, et al. “Grounded language-image pre-training.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Kamath, Aishwarya, et al. “MDETR-modulated detection for end-to-end multi-modal understanding.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[3] Radford, Alec, et al. “Learning transferable visual models from natural language supervision.” International Conference on Machine Learning. PMLR, 2021.

[4] Baldrati, Alberto, et al. “Effective Conditioned and Composed Image Retrieval Combining CLIP-Based Features.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[5] Hendriksen, Mariya, et al. “Extending CLIP for Category-to-image Retrieval in E-commerce.” European Conference on Information Retrieval. Springer, Cham, 2022.

[6] Sérieys, Guillaume, et al. “Text-guided visual representation learning for medical image retrieval systems.” 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022.