WISE Image Search Engine (WISE)

Prasanna Sridhar **Horace Lee** **Abhishek Dutta** **Andrew Zisserman**
{prasanna,horacelee,adutta,az}@robots.ox.ac.uk
Visual Geometry Group (VGG), Dept. of Engineering Science, University of Oxford

Abstract

In order to be useful, a large collection of images needs a search tool that not only searches using image metadata (*e.g*. title, description, etc) but also can search based on image *content*. This paper introduces the WISE Image Search Engine (WISE), an open-source software project based on recent advances in vision-language models that enable content-based image search. Using the images from Wikimedia Commons, we demonstrate the value of a content-based image search tool in retrieving relevant images which would otherwise be "lost" due to missing metadata.

Keywords: content-based image retrieval, WISE

Introduction

Wikimedia Commons is one of the largest repositories of freely usable images on the internet. However, its search tool is based solely on metadata (*e.g*. title, description, *etc*.) which prevents retrieval of images whose metadata is missing or incomplete. The image metadata, if present, often does not fully describe the image content and therefore the existing search tool is unable to retrieve all the images relevant to a search query. This paper introduces the WISE image search engine which leverages recent advances in machine learning and vision-language models that enable search based on image content using natural language. The expressive power of natural language allows the user to flexibly describe what they are looking for. The WISE search latency is comparable to the existing metadata based search tool. The images from Wikimedia Commons are used to demonstrate the benefit of such a visual search engine which can retrieve relevant images that would otherwise be "lost" due to missing metadata.

Background and Related Work

Vision-language models are deep neural networks capable of learning the relationship between concepts described using natural language (*e.g*. a horse drinking water from a river) and those concepts shown in an image (*e.g*. a photograph showing a horse drinking water from a river). Such models are being widely used for generating the caption of an image, retrieving relevant images from a large collection based on image content, rendering an image based on its textual description, *etc*. These models are trained using datasets containing images paired together with text captions/descriptions. Typically, the goal of this training process is to find a vector representation (*i.e*. a sequence of numbers) of images and textual descriptions in a common vector space such that the image and its textual description lie close to each other, while other unrelated image and textual description pairs lie further apart in this high dimensional space. In (Radford et al., 2021), the authors have shown that such a deep model trained on a large dataset consisting of 400 million image-text pairs is capable of learning the relationship between visual concepts and their text description.

Vector-based similarity search is used to find the nearest neighbours of a query feature vector in a high dimensional space. In this work, a pre-trained vision-language model is used to represent both images and text as a point in a common high dimensional (*e.g*. 512) vector space. Therefore, searching a large collection of images becomes a similarity search task which involves finding the points that are closest to the vector representation of the query with respect to some similarity metric.

Our Approach

The WISE software introduced in this paper enables searching over large collection of images using the following two stage approach. In the first stage, each image is represented as a feature vector using a pre-trained vision-language model such as OpenCLIP (Ilharco et al., 2021). The second stage involves building an index of the features such that approximate nearest neighbour search can be performed in the high dimensional feature space. The FAISS (Johnson et al., 2019) open source software is used for the indexing stage. After completion of the feature extraction and indexing stages, the large image collection is ready to be searched using a text query. The search query is represented as a feature vector and the pre-built index is queried to find images in the collection whose vectors have a high cosine similarity with the vector of the search query. These search results are then shown to the user, ranked by their similarity to the query vector.

WISE also supports search using one image or a set of images. An approach similar to text search is applied when the search query is represented by a single image; the only difference is that the query feature vector corresponds to an image instead of textual search query. Such a simplification is possible because vision-language models can project both images and text onto a common feature space. A classifier based search, similar to (Chatfield et al., 2015), is applied if the search query corresponds to a set of images that denote a specific category (*e.g*. Gothic Cathedrals or German Shepherd). A linear binary classifier is trained to distinguish between the set of feature vectors corresponding to the user supplied set of images, which denotes the positive class, and another set of unrelated images that represents a negative class.

Implementation: The capabilities of the WISE search engine are demonstrated based on the images contained in the Wikimedia Commons repository. The Wikimedia Commons data dump created on 2023-01-01 was used to collect the URLs of the JPEG and PNG images in the repository whose size was larger than 224 pixels. We use the img2dataset (Beaumont, 2021) tool to download these images. We use approximate nearest neighbour search method to instantly search millions of images.

Results

We compared the retrieval performance of existing Wikimedia Commons metadata based search engine with the proposed WISE by manually verifying their search results for a set of 9 search queries. The results in Figure 1 show that WISE retrieves more relevant results.

We qualitatively illustrate the performance of these two search engines using two queries. Figure2 shows results for the query "horse near river". The results from Wikimedia contain images that have either a horse or a river but not both in the same image. Figure3 shows results for the query "a cat angry at another cat". The results from Wikimedia contain a cat or images that contain the keywords "cat" or "angry" in the metadata. None of the results from Wikimedia Commons reflect the scene described by the query; the results only match one or two of the words in the query. The results from WISE for both the queries are more relevant.

Since the underlying vision-language model is trained on a large datasets of image-text pairs from the internet, we find that WISE can also retrieve relevant results for search queries about famous people (*e.g*. John Lennon), landmarks (*e.g*. Oxford Radcliffe Camera), etc. without relying on metadata.

Conclusion

This paper described an image search engine based on a vision-language model, and demonstrated the benefits of search based on visual content using images contained in the Wikimedia Commons repository. The WISE software[1] is available as an open source project to encourage adoption of a content-based search capability by organisations, like Wikimedia Commons, who manage large collections of images.

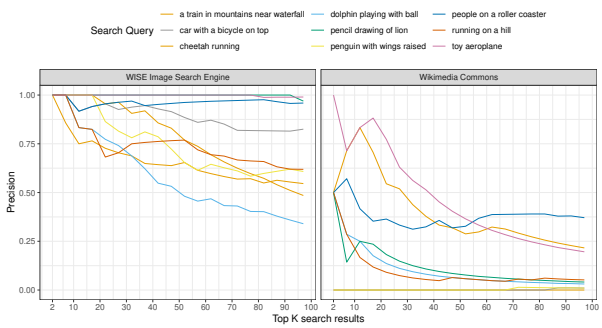

Figure 1: Fraction of relevant instances (*i.e*. precision) among the retrieved image instances (*i.e*. Top K retrieved result) retrieved by the existing Wikimedia Commons metadata based search engine and the proposed WISE Image Search Engine.

Acknowledgement

This work is funded by Visual AI grant (EP/T028572/1).

References

[Beaumont2021] Romain Beaumont. 2021. img2dataset: Easily turn large sets of image urls to an image dataset. `https://github.com/rom1504/img2dataset`.

[Chatfield et al.2015] Ken Chatfield, Relja Arandjelović, Omkar Parkhi, and Andrew Zisserman. 2015. On-the-fly learning for visual search of large-scale image and video datasets. *International Journal of Multimedia Information Retrieval*, 4(2):75–93, June.

[Ilharco et al.2021] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip, July.

[Johnson et al.2019] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

[Radford et al.2021] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[1]Coming soon at `https://www.robots.ox.ac.uk/~vgg/software/wise/`

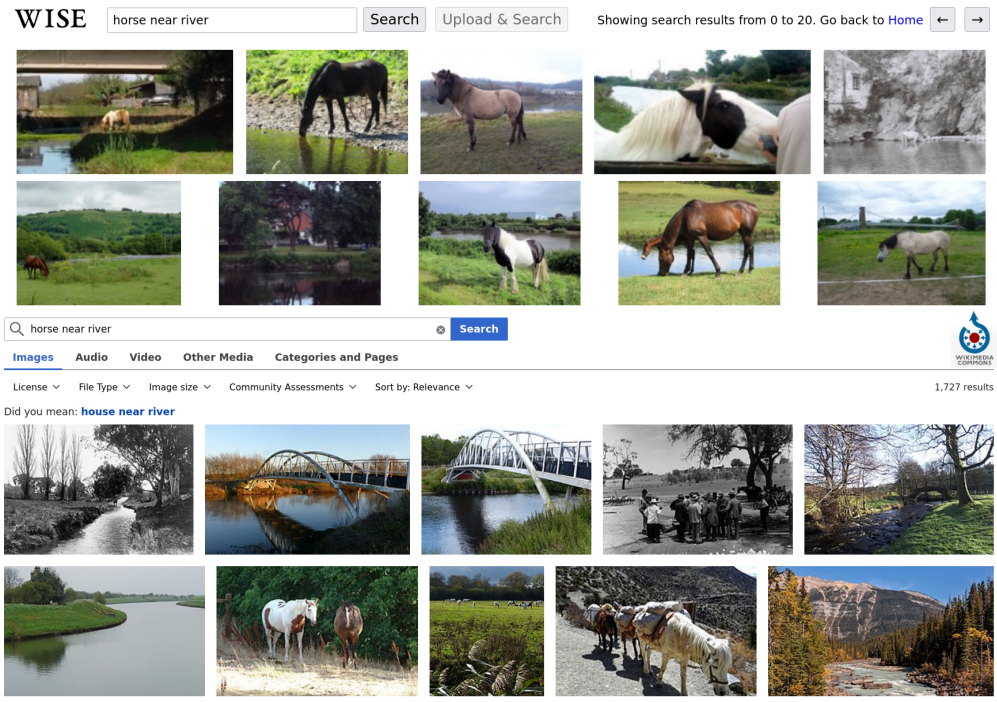

Figure 2: Results for the search query "horse near river" from the WISE image search engine (top) and from the existing Wikimedia Commons' metadata search engine (bottom). The results from the existing search contains images that have either a horse or a river but not both in the same image, while the results from WISE are relevant to the search term as well as confirming the existence of such images in the database.

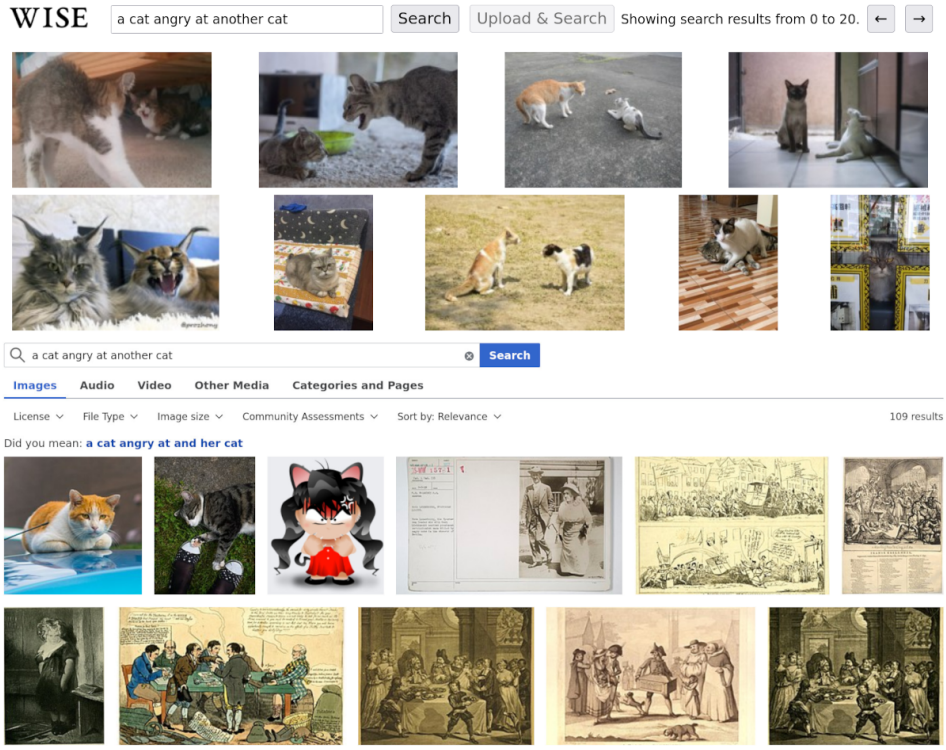

Figure 3: Results for the search query "a cat angry at another cat" from the WISE image search engine (top) and from the existing Wikimedia Commons' metadata search engine (bottom). WISE results show that it can handle complex natural language queries.