QISS: An Open Source Image Similarity Search Engine

Maxime Portaz(✉), Adrien Nivaggioli, Hicham Randrianarivo(✉), Ilyes Kacher, and Sylvain Peyronnet

Qwant Research, Paris, France
{m.portaz,h.randrianarivo}@qwant.com

Abstract. Qwant Image Similarity Search (QISS) is a multi-lingual image similarity search engine based on a dual path neural networks that embed texts and images into a common feature space where they are easily comparable. Our demonstrator, available at http://research.qwant.com/images, allows real-time searches in a database of approximately 100 million images.

Keywords: Neural networks · Image retrieval

1 Introduction

Qwant Image Similarity Search (QISS) is a multi-lingual image search engine. It allows users to make queries both textually or by using images. QISS relies on similarity search. This means that it will compare the content of a query with the data in its index and returns the elements it considers most similar visually or semantically. In our case, we consider the index elements whose Euclidean distance from the query is closest to zero to be the most similar.

If an image and its describing text are close to one another in the representing space, it is possible to query either one with the other. QISS aims to allow the user to use text or image to query a set of images. While other search engines are based on text surrounding images or tags, QISS evaluates the semantic similarity between the query and each element of the database.

In order to process a query, QISS projects it with a Deep Neural Network. QISS is using a dual path Neural Network, that embed different languages and images into on semantic space [5]. It relies on Nvidia TensorRT server[1] for inference. The indexation of roughly 100 millions images, all available through QISS, is done using the Facebook AI Similarity Search (FAISS) library [3]. The QISS project is open source: the code for neural network training is available at https://github.com/QwantResearch/text-image-similarity while the servers that compose QISS are available at https://hub.docker.com. All dockers are accessible at the address https://hub.docker.com/r/<docker_name> and

[1] https://docs.nvidia.com/deeplearning/sdk/tensorrt-inference-server-guide/docs/index.html.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 486–490, 2020.
https://doi.org/10.1007/978-3-030-45442-5_63

can be obtained using the command `docker pull docker_name`. Docker names for this project are: `chicham/text_server`, `chicham/image_server`, `chicham/language_server`, `chicham/index_server` and `chicham/lmdb_server`. We also use nvcr.io/nvidia/tensorrtserver:19.06-py3 as the model server.

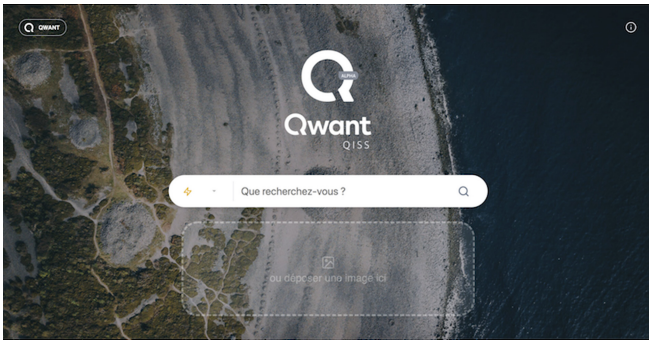

Fig. 1. QISS homepage, where the user can query the index by uploading an image or typing a sentence.

2 System Description

QISS can be used to query the images index by either using texts or images as queries. Also, the representation of the texts is multi-lingual. This means that words from different languages but with similar meanings will have close representations in the semantic space.

2.1 Multi-lingual Text Representation

One of QISS's constraints is to be available in several languages. Instead of using a translation of textual image descriptions, we propose to use multi-lingual word embeddings to cope with multiple languages. Word embeddings are used to project words into a semantic space, where distance and semantic similarity are related. Multi-lingual embeddings, such as Multilingual Unsupervised or Supervised word Embeddings (MUSE) [1], allow for the representation of different languages into one common space. Thanks to this alignment, a neural network can extract information from the embedded words in all learned languages. This allows QISS to have only one index that contains every image, and that can answer to queries expressed into several languages. This is a strong difference with classic search engines that have one index for each language.

2.2 Model for Image and Text Representation

To project both images and texts into the same space, we use two networks trained simultaneously. The image branch of the network uses a Convolutional

Neural Network (CNN) followed by a fully connected layer that embed images. The second branch is a multi-layer Recurrent Neural Network (RNN) that compose a multi-lingual word embeddings list into the same space. This list corresponds to a given sentence.

2.3 Data

We use two datasets to train the models used by QISS. Each dataset is composed of images and their corresponding captions. The first dataset is Common Objects in COntext (COCO) [4]. It contains 123 287 images with 5 English captions per image. The second dataset is called Multi30K [2]. It contains 31 014 images with captions in French, German, and Czech. We use 29 000 images for training and 1014 for validation and 1000 for testing.

MUSE allows for a common representation for 110 languages. Once we trained our model in English using COCO, we used MUSE to transfer the computed embeddings to any language supported by MUSE, at no cost.

For the online demonstration, we indexed images from the Yahoo Flickr Creative Commons (YFCC) [6] image dataset. This dataset contains roughly 100 million images under Creative Commons license.

2.4 Overview

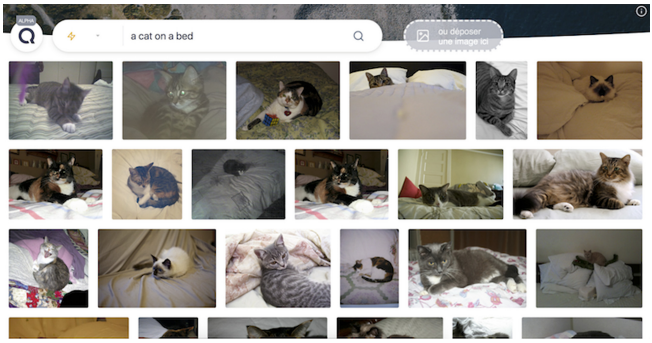

Fig. 2. Result page when the user search for "a cat on a bed", showing the images closest to this text.

As said above, QISS is a full image search engine based on similarity search. Figure 1 shows the interface, where it is possible to search using a text query or by uploading an image. The results are shown in Fig. 2. The images that our method evaluates as the most similar to the query (either text or image) are returned.

The Fig. 3 shows the overview of the system. In our general system, images are taken from a Web Crawler. However, in the context of the online demonstrator research.qwant.com/images, we are using only images from the YFCC dataset. These images go through TensorRT features Extractor, to be then indexed with FAISS.

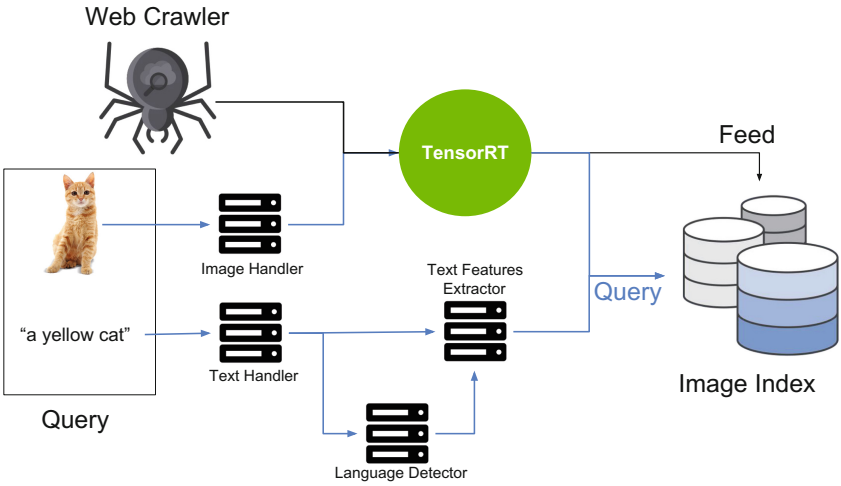

Fig. 3. Overview of the QISS system.

At the query time, the user can:

- Upload an Image. It is sent to the Image Handler and the inference is realized with NVidia TensorRT.
- Search a text. The text goes to the language detector and the Text Features Extractor.

References

1. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: International Conference on Learning Representations (2018). https://doi.org/10.1111/j.1540-4560.2007.00543.x. http://arxiv.org/abs/1710.04087
2. Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30K: multilingual English-German image descriptions. In: Proceedings of the 5th Workshop on Vision and Language. Association for Computational Linguistics, Stroudsburg (2016). https://doi.org/10.18653/v1/W16-3210. http://arxiv.org/abs/1605.00459. http://aclweb.org/anthology/W16-3210
3. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734 (2017)

4. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48. http://arxiv.org/abs/1405.0312
5. Portaz, M., Randrianarivo, H., Nivaggioli, A., Maudet, E., Servan, C., Peyronnet, S.: Image search using multilingual texts: a cross-modal learning approach between image and text. Ph.D. thesis, Qwant Research (2019)
6. Shamma, D.A.: One hundred million creative commons Flickr images for research, 24 June (2014)