clarifai

Machine Learning for Visual Similarity Search

Introduction

Early search engines powered by the likes of Yahoo! and Alta Vista offered simple keyword matching technology that helped users find content on the web. These services would count up the number of times a given search term was present on a web page and then rank search results based on keyword frequency. This approach was later augmented by Google and others where search results relevance were improved by analysing the relationship between different websites.

With advancements in machine learning, new techniques have been developed that allow users to search for content without using keywords at all (Yang et al. 2017). This paper explores how machine learning has been applied to image processing to enable search systems that can use images instead of search terms and return search results based on visual similarity alone.

Practical applications

There are several practical applications of visual and text similarity search. One can authenticate the images, find the source of the image, and authenticate identities through facial recognition and face matching. In E-commerce, visual search can be used to allow customers to search for a product using an image. Retailers can also offer product recommendations based on visual search preferences. Another application used by entertainment companies includes using pictures of celebrities to search for their names online. Visual search is also used in education where students can upload images or videos without descriptions and uncover similar images. (Biggs and Mitroff 2019). Text similarity search can be used in chatbot services, or to research similar articles. It can also be used in the analysis of image captions.

Machine learning approaches to visual search

Most research in the field of visual search has used Convolutional Neural Networks (CNN) for classifying the images as well as image retrieval (Zhang et al. 2018). The use of CNN resulted in superior visual research and more accurate matching of the images when compared to traditional methods such as the Fisher Vector and Vector of Locally Aggregated Descriptors (VLAD). CNN can take an image and find the best possible matches as shown in the following figure:

Figure 1: CNN based visual search

A similar approach was adopted in another study where features were extracted from the OverFeat network as a generic image representation (Zhong et al. 2017). This approach used the cropped and rotated image samples. For each image, multiple sub-patches of different sizes were obtained. The distance between the reference and query image was computed for each sub-patch using CNN. The lesser distance meant a more similar image and vice versa. The problem with this and other similar techniques was patching, cropping, and rotating images in different directions took significant time. The solution was to use holistic descriptors where the whole image is mapped to a single vector with the CNN model. The results show that the principal component analysis was less affected when compared to Fisher vectors and VLAD.

How CNN is used in Visual Search

In the first step, the CNN are used for feature extraction from the image. A CNN image classifier is used to convert images to low dimensional feature vectors representing the "features" learned by the network. A trained CNN model can also be used by removing its last high-level layers that were used to classify the objects and use a dissected model to convert the input image into feature vectors as shown below:

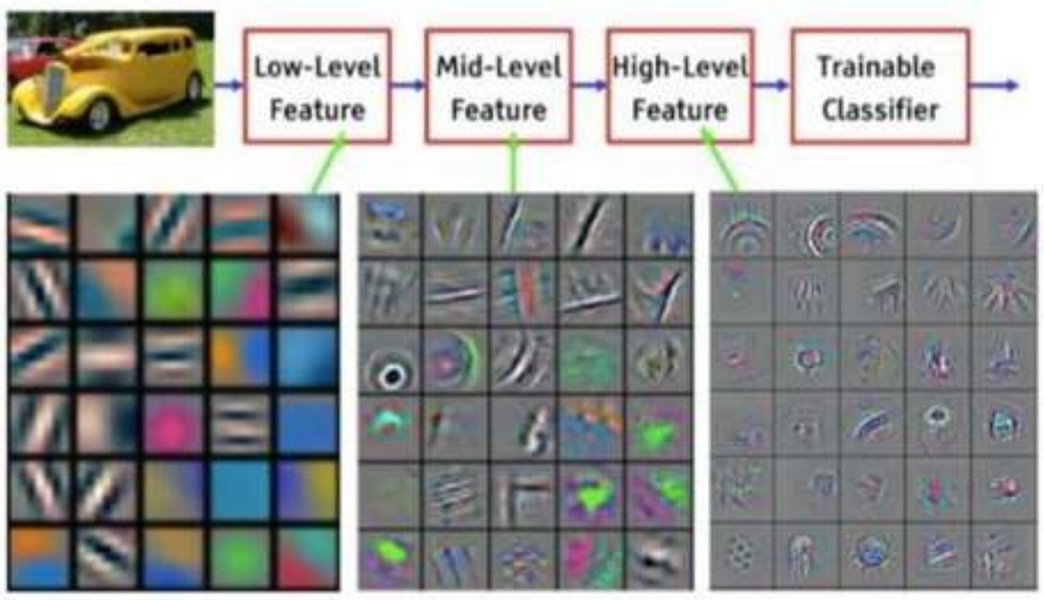

Figure 2: CNN feature vector extraction

These feature vectors tend to contain noise and redundant information. In order to make the image retrieval process and filtering the most important information efficient, the data is further compressed based on the principal component analysis. The 2048 dimensions are reduced to 256 dimensions as shown below:

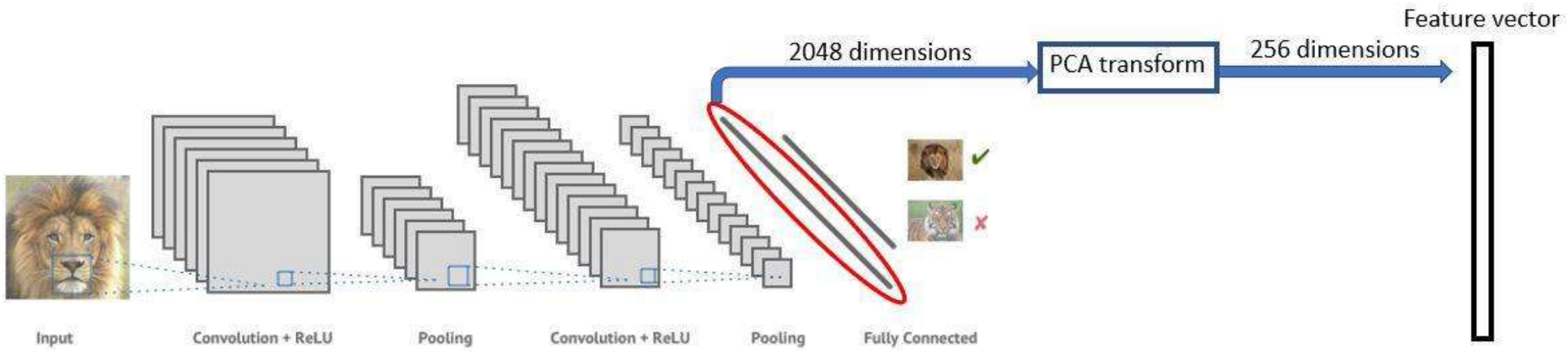

Figure 3: Noise filtering in the feature vector

Now a noise-free/filtered feature vector is obtained. The second step in visual research is using a similarity matrix to compare the images. A variety of different matrices can be used but one of the most popular is the “cosine similarity” matrix. This method measures the angle between images in high dimensional feature space. People cannot visualize 256 dimensional space, so a simple two-dimensional plane is shown here for demonstration purposes. It can be observed that for the two similar images, angle ϕA is less.

The image B contains different objects and angle ϕB is also large.

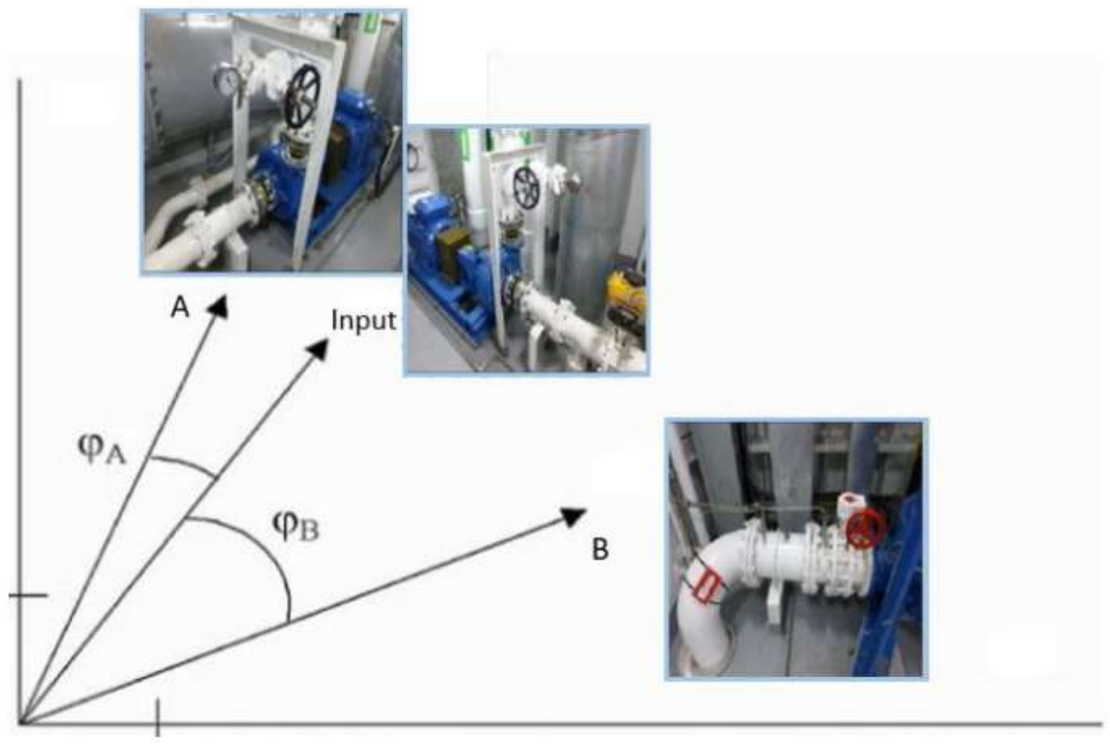

Figure 4: Cosine similarity matrix for visual search

Another common technique to measure the similarity between the two images is using image clusters. In this method, the initial image is passed through the feature extractor to create the feature vector, as in the previous step. This feature vector is a quantitative description in terms of image content. This information can be passed through a clustering algorithm which will relocate the images to respective clusters of similar images. During this process, it may not be clear which individual images are present in a specific cluster, however, it's certain that the images within the various clusters will contain similar content.

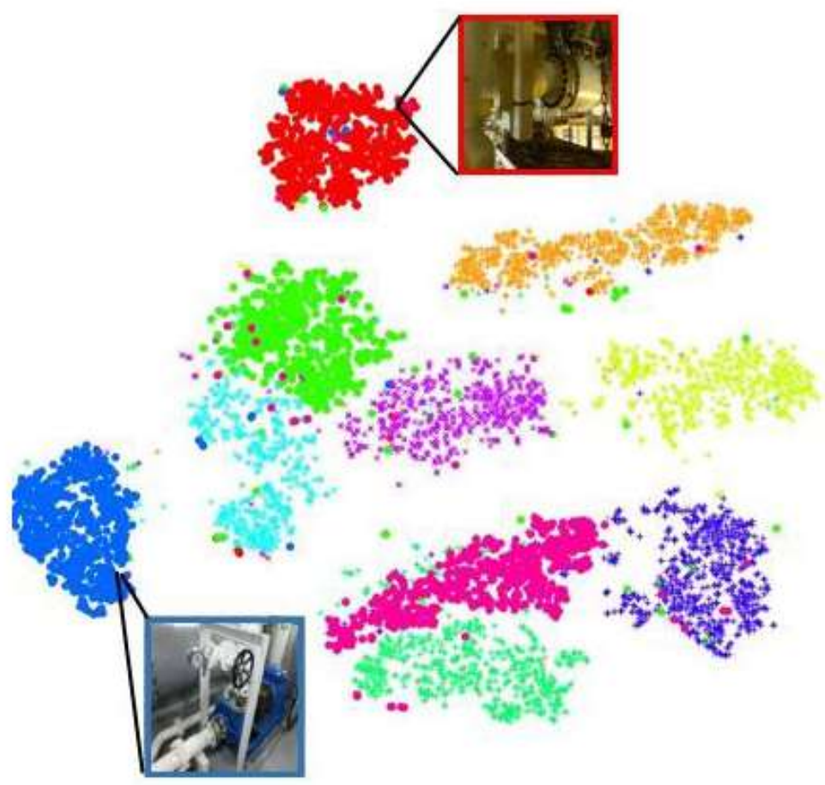

Figure 5: Feature vector to clustering for visual search

Machine learning approaches to text similarity search

Clustering and cosine similarity techniques are also used when comparing text passages for similarity. But in the case of text similarity, the feature vectors analyze word and sentence structure. By using machine learning approaches, entire passages of text can be compared to one another, and the order of words can be taken into account when determining the meaning of a passage.

This allows for a more robust notion of similarity than can be achieved with traditional approaches like ranking documents based on how many words they share with the query. A document may be similar to the query even if they have very few words in common.

Conclusion

In summary, while visual similarity search techniques are still evolving, comparing the cosine similarity of two images, or using clustering techniques to group images together have proven effective in a variety of scenarios. As machine learning algorithms evolve and data labeling services improve over time, the features identified by CNN will become more accurate and less noisy — improving the performance of the visual search. AI platforms like Clarifai could be used to provide visual similarity search without any special configuration and tooling.

About Clarifai

Clarifai is the leading independent provider of deep learning AI for unstructured data. Headquartered in NYC, the company was founded in 2013 by Matt Zeiler, Ph.D. after winning the top 5 places at ImageNet with the goal of delivering state-of-the-art AI technologies to enterprises and organizations around the world. Clarifai offers the most powerful platform for the end-to-end AI lifecycle, UIs that unleash deep learning for every skill set, and the best solutions to important use cases. The company has raised $40M from Union Square Ventures, Menlo Ventures, Lux Capital, NVIDIA, Google Ventures, and Qualcomm. Clarifai continues to grow with 100+ employees at its headquarters in New York City, and offices in San Francisco, Washington, DC, and Tallinn, Estonia.

Sources:

Biggs, Adam T, and Stephen R Mitroff. 2019.
“Visual Search Training via a Consistency Protocol: A Pilot Study.”
Visual Cognition 27 (9–10): 657–67.

Yang, Fan, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. 2017. “Visual Search at Ebay.”
In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2101–10.

Zhang, Yanhao, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. 2018. “Visual Search at Alibaba.”
In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 993–1001.

Zhong, Chunlin, Yi Yu, Suhua Tang, Shin’ichi Satoh, and Kai Xing. 2017.
“Deep Multi-Label Hashing for Large-Scale Visual Search Based on Semantic Graph.”
In Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data, 169–84. Springer.