DATA-643 Assignment - 06

Mohamed Elmoudni
Shazia Khan
Senthil Dhanapal

Contents

Introduction

The innovative system we will be analyzing is called the similarity search on Flickr. It is photo based search based on color feature and style. it is also called Semantic Similarity search - a search based on photos.

Analysis

Semantic similarity search is done by using neural network method:- vectors go through several transformation until a fixed constraint. The output vector or feature vector consists of several thousands of dimensions. Then few similarity algorithms are being applied on the output vectors to determine the photos similarity. The most common similarity algorithms are Euclidean distance, called Locally Optimized Product Optimization (LOPQ).

the Euclidean distance group similar photos and well suited for small datasets. However, matching vectors against a large number of photos, billions of images, the performance degrades and the Euclidian similarity search becomes very slow and comes to a halt.

To overcome the Euclidean distance performance limitations, Flickr issues approximate nearest algorithm called Locally Optimized Product Optimization (LOPQ). The LOPQ clusters index vectors uses k-means clustering while maintaining index-clusterid pairs. Then, during querying, the LOPQ finds the query vector's clusterid and all the vectors within that cluster. For instance, For 1 billion photos, 1 million clusters required to store 1000 photos per cluster. Querying requires matching 1 million clusters to find nearest cluster.

In addition, to further improve performance, vectors are broken into sub-vectors and each sub-vector is assigned a cluster, which will reduce the cluster matching by the number of times of sub-vectors count. The idea of breaking vectors into sub-vectors and assigning each sub-vector a cluster is known as product quantization. Using this idea to index a dataset is known as inverted multi-index.

Use case

Below is an example how user can leverage the new way of searching photos..

Member Experience of using Flickr Similarity Search We logged in to Flickr as member searched for ‘shoes’.

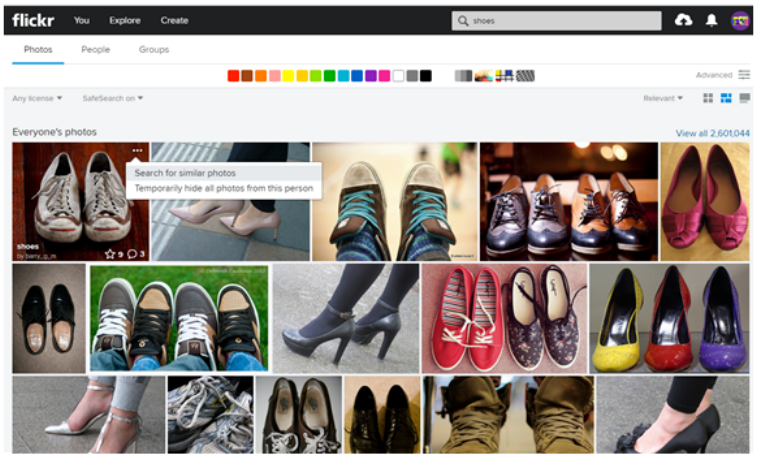

Figure 1:

We selected the first shoe picture and clicked on “Search for similar photos”. Our expectations were to find some more of worn-out sneakers in black and subdued colors. However, we were mistaken. The “Search for similar photos”, resulted in worn-out sneakers but also picked up unexpected results, such as the picture of a man and a woman in black and white muted colors.

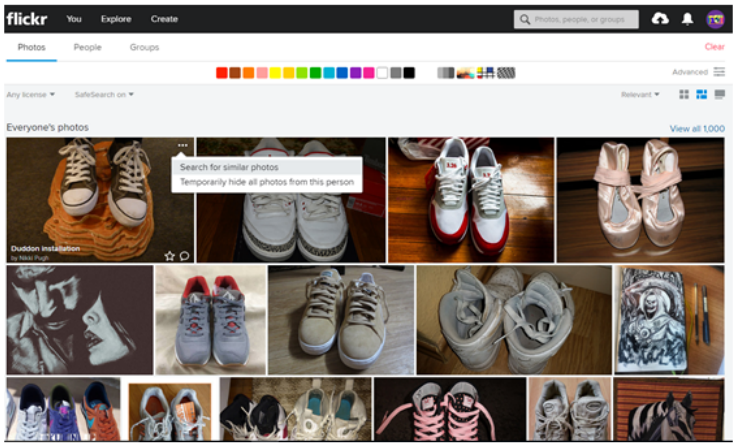

Figure 2:

Again clicking on the first picture's "Search for similar photos", we found more shoes and also pictures of an arm with tattoo. Our understanding is that the Neural Network used in the Search Code is looking at the overall similarities in shape and colors of the shoe and tattooed arm as seen in the picture.

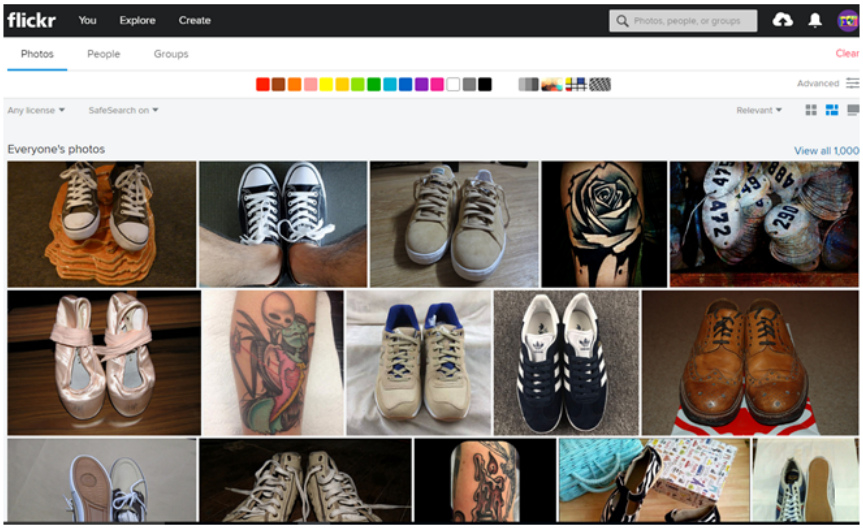

Figure 3: