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

Flickr uses Semantic Similarity search:- search based on semantic of photos. Semantic similarity search is done by using neural network method where vectors go through several transformation until a fixed constraint. Output vectors or feature vectors consist of several thousands of dimensions. Using Euclidean distance vectors are grouped into multiple similar groups. However, Storage of high deminsional data of billions of images into clusters is still unmanagable and querying for a matching vector on such a large index becomes an expensive process. To overcome theses issues flickr uses approixmate nearest algorithm called Locally Optimized Product Optimization (LOPQ). LOPQ clusters index vectors using k-means clustering and maintain index-clusterid pair. While querying for a vector, flickr uses query vector's clusterid to find all vectors within that cluster. Still for 1 billion photos, 1 million clusters required to store 1000 photos per cluster. Querying requires matching all 1 million clusters to find nearest cluster. This still leads to performance issue. To handle that further, flickr algorithm breaks vectors into subvectors and each subvector is assigned a cluster, which will reduce the number of clusters to match by the number of subvectors count. The idea of breaking vectors into subvectors and assigning each subvector a cluster is known as product quantization. Using this idea to index a dataset is known as inverted multi-index. Next step it uses is product quantization on the residuals of data to rank the set of candidates. The residual of a point is the difference vector between the point and its closest cluster centroid. Instead of storing the residuals, LOPQ product quantizes the residuals, usually with a higher number of splits, and stores only the cluster indexes in the index. Split vectors and centroid info can be stored in 8 bytes but with some loss of info. Vector can be reconstructed using quantization code and looking up corresponding centroid and concatenate all the centroids for this vector and it can also approximate the distance from the query to an index vector by computing the distance between the query and the reconstructed vector. This helps faster computation for many candidate points by computing the squared difference of each split of the query to all of the centroids for that split. After computing this table, it computes the squared difference for an point by looking up the precomputed squared difference for each of the indexes and summing them together to get the total squared difference. This trick allows to quickly rank many candidates without resorting to distance computations in the original vector space. LOPQ is state-of-the-art for quantization methods, and one can find more information about the algorithm, as well as benchmarks in http://image.ntua.gr/iva/research/lopq/

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’.

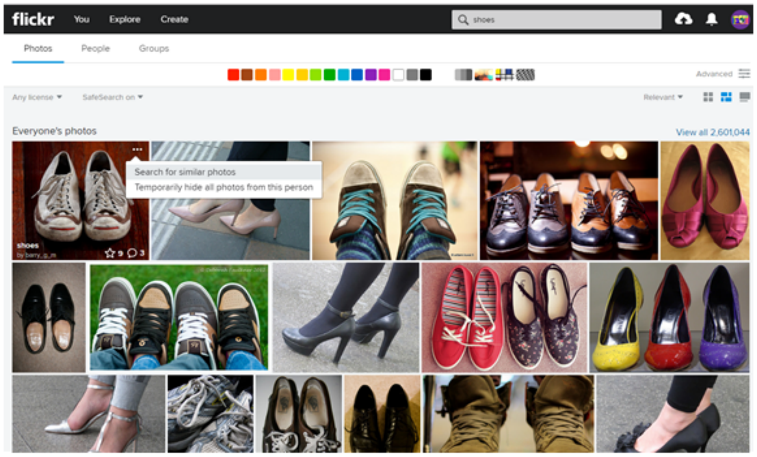

Figure 1:

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.

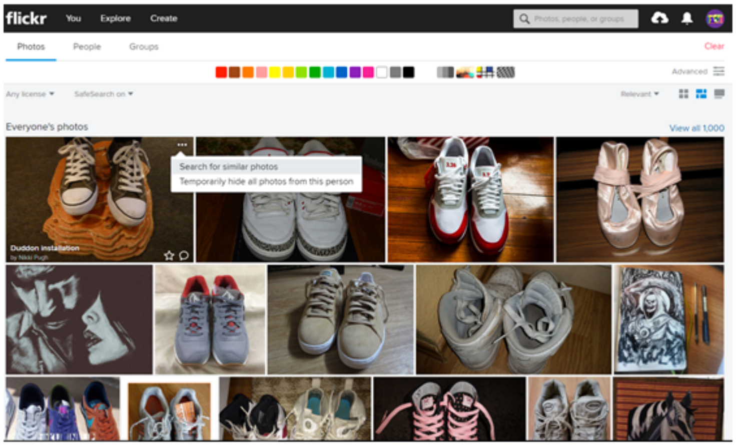

Figure 2:

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.

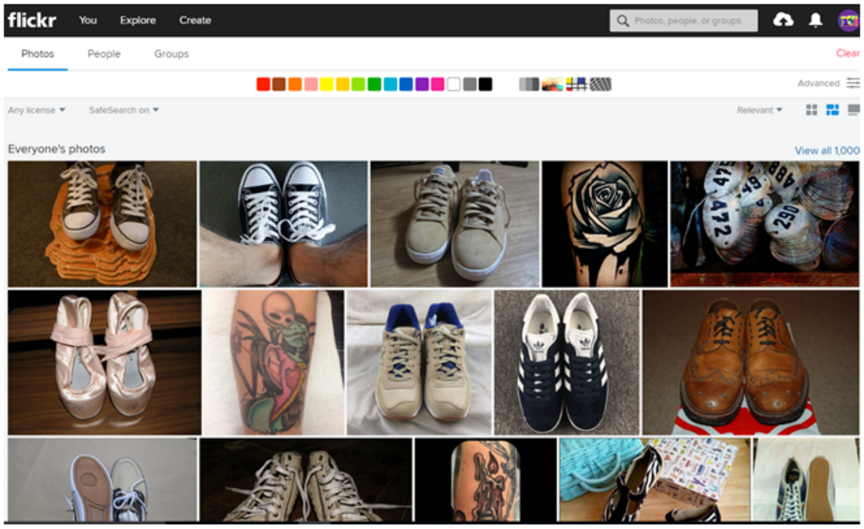

Figure 3: