1 Evaluation metrics

The main goal of evaluating the model's performance of a CBIR system is to measure the quality of the image retrieval for a given query image. Two non-overlapping image sets *Gallery*, *Query*, and two widely used metrics mAP and nDCG are chosen to measure the retrieval performance. After the training process, for every gallery images, X_i are converted to hash code h_i. The hash code h_j is obtained from given an image X_j from the *Query* set. The relation between cosine similarity and normalized Euclidean distance for two hash codes h_i and h_j with length K is defined by,

$$d(h_i, h_j) = \frac{K}{4} \left\| \frac{h_i}{\|h_i\|} - \frac{h_j}{\|h_j\|} \right\|_2^2 = \frac{K}{2}(1 - cos(h_i, h_j)) \tag{1}$$

The discussion of the computation process for both two metrics is given in the following subsections.

1.1 Mean Average Precision

mAP, which has been established to have very strong discrimination and stability, is utilized to measure the ranking quality of image retrieval from the *Gallery* in our tests to assess performance. We are ranking the *Gallery* set $\{X_i\}_i$ with respect to the Hamming distances in descending order and the collected top retrieval images. For topmost p retrievals, mAP values are represented as $mAP@p$. $mAP@p$ is calculated based on the average precision (AP) $AP@p$ for $top - p$ retrievals. After retrieving the relevant images from the *Gallery* with respect to a query image, the set of $top - p$ retrievals is employed to obtain the AP. Let X_m is m-th ranked images from *Gallery* for query image X_j in *Query* set then,

$$AP_j@p = \frac{\sum_{m=1}^{p} Prec_j(m) Rel_j(m)}{\sum_{m=1}^{p} Rel_j(m)} \tag{2}$$

Next, the precision value for $top - m$ retrievals of *query* image X_j is denoted by $Prec_j(m)$ and defined by,

$$Prec_j(m) = \frac{\sum_{m=1}^{m} Rel_j(m)}{m} \tag{3}$$

where, $Rel_j(m) = 1$, if $(X_m, X_j) \in type\ 1 \cup type\ 2$, other wise $Rel_j(m) = 0$.

Finally, the value of $mAP@p$ is the mean of $AP_j@p \forall X_j \in \{Query\}$.

1.2 Normalized Discounted Cumulative Gain

The nDCG metric is most commonly employed in information retrieval tasks like ranking the effectiveness of a search algorithm or related applications. To compute $nDCG@p$ for $top - p$ retrieval, first we need to compute Cumulative Gain ($CG@p$), and Discounted Cumulative Gain ($DCG@p$) for $top - p$ retrieval.

Let X_m is m-th ranked images from *Gallery* for query image $X_j \in \{Query\}$ then, The relevance for image X_m is defined by,

$$Rel_j(m) = \begin{cases} 2, & \text{if } (X_m, X_j) \in type\ 1 \\ 1, & \text{if } (X_m, X_j) \in type\ 2 \\ 0, & \text{if } (X_m, X_j) \in type\ 3 \end{cases}$$

The mathematical formulation for $DCG@p$ is given by,

$$DCG_j@p = \sum_{m=1}^{p} \frac{2^{Rel_j(m)} - 1}{log_2(m+1)} \tag{4}$$

$DCG@p$ becomes a larger list for all query images from the *Query* set. We normalize this by dividing maximally achievable value or Ideal DCG (iDCG). Finally,

$$nDCG@p = \frac{DCG@p}{iDCG@p} \tag{5}$$

Where iDCG@p = DCG@p of ideal ranking.