Learning with Average Precision: Training Image Retrieval with a Listwise Loss

Jérôme Revaud  Jon Almazán  Rafael S. Rezende  César Roberto de Souza
NAVER LABS Europe

1. Backpropagation of mAPq

In this section, we expand upon Section 3.2 of our submission to further explain the backpropagation process for a batch of image descriptors \( D = [d_1, \ldots, d_B] \) and groundtruth \( Y = [Y_1, \ldots, Y_B] \). Additionally, we will demonstrate the differentiability of the loss function \( L \). We separate the algorithm in three stages, accordingly to Fig. 3 of our submission.

**Stage 1.** This stage consists only of a forward pass of our network in evaluation mode, i.e., no backpropagation is needed. For each image \( I_i \) of the batch, its descriptor \( d_i \) is computed as \( d_i = f_\theta(I_i) \) and the matrix \( D \) is determined by the stacking of these descriptors by the rows.

**Stage 2.** In this stage, we compute the loss \( \ell = L(D, Y) \) given by:

\[
\ell = 1 - \frac{1}{B} \sum_{q=1}^{B} \sum_{m=1}^{M} \hat{P}_m(d_q^\top D, Y_q) \Delta \hat{r}_m(d_q^\top D, Y_q). \tag{1}
\]

Since, at this stage, the backpropagation method stops at the descriptors level, our goal is to compute the derivatives \( \frac{\partial \ell}{\partial d} = \left( \frac{\partial \ell}{\partial d_i} \right)_{1 \leq i \leq B} \). We introduce some mathematical notations to ease the readability of this section. Let \( S^q = d^\top D \) in \([-1, 1]^B \) denote the similarities of the batch images with a query \( I_q \) in the batch, \( \hat{\delta}_m = \delta(S^q, m) \) in \([0,1]^B \) denote the soft-assignment of \( S^q \) to the \( m \)-th bin and \( \xi_m = \sum_{k=1}^{m} \delta_k \) in \( \mathbb{R}_+^B \) denotes the sum of all the soft-assignments of \( S^q \) up to the \( m \)-th bin. Moreover, \( \hat{P}_m = \hat{P}_m(S^q, Y_q) \) and \( \Delta \hat{r}_m = \Delta \hat{r}_m(S^q, Y_q) \) denote the precision and incremental recall at bin \( m \) of \( S^q \), respectively. These notations allow us to rewrite \( \hat{P}_m \) and \( \Delta \hat{r}_m \) as:

\[
\hat{P}_m = \left( \xi_m \right)^\top Y_q \left( \xi_m \right)^\top \mathbf{1} \tag{2}
\]

\[
\Delta \hat{r}_m = \frac{\left( \delta_m \right)^\top Y_q}{N^q}. \tag{3}
\]

We now derive Eq. (1) w.r.t. \( d_i \) and, by applying the chain-rule, we obtain:

\[
\frac{\partial \ell}{\partial d_i} = -\frac{1}{B} \sum_{q=1}^{B} \sum_{m=1}^{M} \left( \frac{\partial \hat{P}_m}{\partial d_i} \Delta \hat{r}_m + \hat{P}_m \frac{\partial \Delta \hat{r}_m}{\partial d_i} \right) \tag{4}
\]

\[
= -\frac{1}{B} \sum_{q=1}^{B} \sum_{m=1}^{M} \left( \Delta \hat{r}_m \frac{\partial \hat{P}_m}{\partial \xi_m} \frac{\partial \xi_m}{\partial S^q} + \hat{P}_m \frac{\partial \Delta \hat{r}_m}{\partial \xi_m} \frac{\partial \xi_m}{\partial S^q} \right) \frac{\partial S^q}{\partial d_i}. \tag{5}
\]

We can separately compute each of the terms of Eq. (5). The derivatives \( \frac{\partial \hat{\delta}_m}{\partial S^q} \) and \( \frac{\partial \xi_m}{\partial S^q} \) in \( \mathbb{R}^{B \times B} \) are defined and can be backpropagated thanks to the definition of the triangular kernel:

\[
\left( \frac{\partial \xi_m}{\partial S^q} \right)_{ij} = \left( \sum_{k=1}^{m} \frac{\partial \delta_k}{\partial S^q} \right)_{ij}, \tag{6}
\]

\[
\left( \frac{\partial \hat{\delta}_m}{\partial S^q} \right)_{ij} = -\frac{\text{sign}(S^q - b_m)}{\Delta} \mathbb{I} \left[ |S^q - b_m| \leq \Delta \right] \mathbb{I}[i = j]. \tag{7}
\]

Next, the missing derivatives, \( \frac{\partial \hat{P}_m}{\partial \xi_m} \) and \( \frac{\partial \Delta \hat{r}_m}{\partial \xi_m} \) in \( \mathbb{R}^{1 \times B} \) and \( \frac{\partial S^q}{\partial d_i} \) in \( \mathbb{R}^{B \times D} \) can be expressed as follows:

\[
\frac{\partial \hat{P}_m}{\partial \xi_m} = \frac{1}{(\xi_m^\top \mathbf{1})^2} \left( \xi_m^\top \mathbf{1} Y_q^\top - Y_q \mathbf{1}^\top \right), \tag{8}
\]

\[
\frac{\partial \Delta \hat{r}_m}{\partial \xi_m} = \frac{1}{N^q} Y_q^\top, \tag{9}
\]

\[
\left( \frac{\partial S^q}{\partial d_i} \right)_{j} = d_j^\top \mathbb{I}[i = q] + d_q^\top \mathbb{I}[i = j]. \tag{10}
\]

And finally, putting together Eqs. (6), (7), (8), (9), (10) in Eq. (5) proves that \( L \) is differentiable and its derivative provides useful information for backpropagation.
Stage 3. In the last stage, the goal is to compute $\frac{\partial \ell}{\partial \Theta}$. For each $i$ in $[1, \ldots, B]$, $d_i = f_\Theta(I_i)$ and $\frac{\partial d_i}{\partial \Theta} = \frac{\partial f_\Theta(I_i)}{\partial \Theta}$ are computed according to the network backpropagation algorithm, and $\frac{\partial \ell}{\partial \Theta}$ is obtained in stage 2. Therefore, the chain-rule gives us:

$$\frac{\partial \ell}{\partial \Theta} = \sum_{i=1}^{B} \frac{\partial \ell}{\partial d_i} \frac{\partial d_i}{\partial \Theta}$$.

2. APQ is a good approximation of the AP

In this section, we illustrate the quality of the approximation given by APQ to the true AP. In order to assess this approximation, we scatter plot the true AP and APQ for all queries of the \textit{T}Paris dataset in Fig. 1. We observe a correlation of ~0.98 between them.

![Figure 1. Scatter plot of the true AP versus APQ for \textit{T}Paris [2].](image)

3. Hyperparameters

In this section, we present additional information regarding the Table 1 in our original submission. In Table 1 above we give an explicit list of the hyperparameters used by each method. As it can be seen, our approach requires much fewer hyperparameters than [1, 3]. This means much less effort is required during training since there are less variables to tune and to optimize using costly procedures such as grid-searches and cross-validation.

Specifically, in order to use hard negative mining (HNM) as in [1], one would need to find correct values for at least 6 hyperparameters: the number $N$ of images in mining pool; the $n$ for when selecting the top-$n$ hardest triplets for mining; the number $k$ of times before updating good candidates; the number $b$s of triplets before an update; the margin $m$ of the loss; the batch size $b$. In contrast, our loss only requires 2 hyper-parameters which have safe defaults that are not very sensitive to changes (please refer to Figures 4 and 5 in our submission).

### Abbreviations:
- $N$, images in mining pool;
- $n$, top-$n$ hardest triplets for hard negative mining (HNM);
- $k$, number of times before updating good candidates;
- $b$, number of triplets before an update;
- $m$, scale-change threshold;
- $n_h$, number of times hard negatives are re-mined;
- $n_m$, number of negative images in the tuple; $n_b$, number of bins.

4. Multistaged backpropagation algorithm

In this section, we provide in Algorithm 1 a pseudocode for our multistaged backpropagation, as presented in Section 3.3 of our submission. The code is split similarly to Figure 2 of our submission in order to make it easier to follow.

### Algorithm 1: Efficient training for any batch size.

Input: Training batch $B = \{(I_1, y_1), \ldots, (I_B, y_B)\}$

begin

```
# Stage 1
for i in 1..B; do
    compute $d_i = f_\Theta(I_i)$
end

# Stage 2
create $D = [d_1, \ldots, d_B] \in \mathbb{R}^{B \times C}$
create $Y \in \{0, 1\}^{B \times B}$ such that $Y_{ij} = \mathbb{I}[y_i = y_j]$
\ell \leftarrow L(D, Y) \quad \# \text{compute the loss}
compute $\frac{\partial \ell}{\partial Y} \in \mathbb{R}^{B \times C}$

# Stage 3
init $\frac{\partial \ell}{\partial \Theta} \leftarrow 0$
for i in 1..B; do
    compute $d_i = f_\Theta(I_i)$
    compute $\frac{\partial d_i}{\partial \Theta} \leftarrow \frac{\partial f_\Theta(I_i)}{\partial \Theta}$
    accumulate $\frac{\partial \ell}{\partial \Theta} \leftarrow \frac{\partial \ell}{\partial \Theta} + \frac{\partial \ell}{\partial Y} \left( \frac{\partial Y}{\partial \Theta} \right)_i \quad \# \text{w/chain-rule}
end
update $\Theta \leftarrow \Theta - \gamma \frac{\partial \ell}{\partial \Theta} \quad \# \text{optimizer step}
```

end
5. Qualitative results

In this section, we provide additional examples of queries and their respective top-10 result images as obtained by our retrieval model (referred as “GeM (AP)” in Table 1 of our submission) in the $\mathcal{R}$Paris (Table 2) and $\mathcal{R}$Oxford (Table ??) datasets [2]. In these tables, each retrieved image is marked with a colored border indicating whether they belong to the positive or negative groups of the medium benchmark associated with these datasets (see Section 4.1 of our submission).

Failure cases. At the bottom of these tables, we also list the three worst performing queries for each dataset. Those queries serve as demonstrations of failure cases that can be used to indicate in which situations our models could be improved. As it can be seen, correctly retrieving images belonging to these queries requires a significant level of attention to details, such as subtle differences in adornment statues (see last rows of Tables 2 and ??) or similarly looking building façades seen from a short distance (see last rows of Table ??).

Hard benchmark. We also include two other tables (Tables ?? and ??) with exactly the same queries as Tables 2 and ??, but showing their evaluation under the hard benchmark of $\mathcal{R}$Paris and $\mathcal{R}$Oxford. In this protocol, we remove easy-to-retrieve images, focusing instead on difficult cases. While this results in a higher amount of mistakes, our model still manages to correctly retrieve images with varying lighting conditions (see Table ??) and high levels of occlusion (see 2nd row of Table ?? and 5th row of Table ??).

References


Table 2. Top-10 results for 5 random queries and 3 worst-performing queries in \( \mathcal{R}_{\text{Paris}} \). [2]

<table>
<thead>
<tr>
<th>Query</th>
<th>Top-1</th>
<th>Top-2</th>
<th>Top-3</th>
<th>Top-4</th>
<th>Top-5</th>
<th>Top-6</th>
<th>Top-7</th>
<th>Top-8</th>
<th>Top-9</th>
<th>Top-10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random queries</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image1" alt="Image 1" /></td>
<td><img src="image2" alt="Image 2" /></td>
<td><img src="image3" alt="Image 3" /></td>
<td><img src="image4" alt="Image 4" /></td>
<td><img src="image5" alt="Image 5" /></td>
<td><img src="image6" alt="Image 6" /></td>
<td><img src="image7" alt="Image 7" /></td>
<td><img src="image8" alt="Image 8" /></td>
<td><img src="image9" alt="Image 9" /></td>
<td><img src="image10" alt="Image 10" /></td>
<td><img src="image11" alt="Image 11" /></td>
</tr>
<tr>
<td><strong>Top failure cases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image12" alt="Image 12" /></td>
<td><img src="image13" alt="Image 13" /></td>
<td><img src="image14" alt="Image 14" /></td>
<td><img src="image15" alt="Image 15" /></td>
<td><img src="image16" alt="Image 16" /></td>
<td><img src="image17" alt="Image 17" /></td>
<td><img src="image18" alt="Image 18" /></td>
<td><img src="image19" alt="Image 19" /></td>
<td><img src="image20" alt="Image 20" /></td>
<td><img src="image21" alt="Image 21" /></td>
<td><img src="image22" alt="Image 22" /></td>
</tr>
</tbody>
</table>