Vector of Locally Aggregated Descriptors (VLAD) [29] is a popular descriptor pooling method for both instance level retrieval [29] and image classification [23]. It captures information about the statistics of local descriptors aggregated over the image. Whereas bag-of-visual-words [15, 73] aggregation keeps counts of visual words, VLAD stores the sum of residuals (difference vector between the descriptor and its corresponding cluster centre) for each visual word.

Formally, given N D-dimensional local image descriptors $\{\mathbf{x}_i\}$ as input, and K cluster centres ("visual words") $\{\mathbf{c}_k\}$ as VLAD parameters, the output VLAD image representation V is $K \times D$-dimensional. For convenience we will write V as a $K \times D$ matrix, but this matrix is converted into a vector and, after normalization, used as the image representation. The (j, k) element of V is computed as follows:

$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i) \left(x_i(j) - c_k(j)\right), \qquad (1)$$

where $x_i(j)$ and $c_k(j)$ are the j-th dimensions of the i-th descriptor and k-th cluster centre, respectively. $a_k(\mathbf{x}_i)$ denotes the membership of the descriptor $\mathbf{x}_i$ to k-th visual word, *i.e.* it is 1 if cluster $\mathbf{c}_k$ is the closest cluster to descriptor $\mathbf{x}_i$ and 0 otherwise. Intuitively, each D-dimensional column k of V records the sum of residuals $(\mathbf{x}_i - \mathbf{c}_k)$ of descriptors which are assigned to cluster $\mathbf{c}_k$. The matrix V is then L2-normalized column-wise (intra-normalization [4]), converted into a vector, and finally L2-normalized in its entirety [29].