HOW DOES OUR VLAD MODULE WORKS?

Inspired by PointNetVLAD[1] and NetVLAD[2], we can indirectly describe the high-level semantic feature by the relationship of each point's low-level geometric descriptor with a few visual words. The presented PointNetVLAD leverages on the success of PointNet and NetVLAD to do 3D point cloud based retrieval for large-scale place recognition. As shown in Fig. (a), given 6 2-dimensional points $\{p_i | i \in [1,6]\}$ as input, and 4 cluster centres ("visual words") $\{C_k | k \in [1,4]\}$ as VLAD parameters, the output VLAD image representation $\boldsymbol{V}$ is a 4×2 matrix. The (j, k) element of $\boldsymbol{V}$ is computed as follows:

$$V(j,k) = \sum_{i=1}^{6} a_k(p_i)(p_i(d) - c_k(d)) \qquad d \in [1,2]$$

where $p_i(d)$ and $c_k(d)$ are the d-th dimension of the i-th point and k-th cluster center, respectively. $a_k(p_i)$ denotes the membership of the descriptor p_i to k-th visual word, i.e. it is 1 if cluster c_k is the closest cluster to descriptor p_i and 0 otherwise. Thus, the PointNetVLAD is designed to aggregate local features into the visual words. In this way, the VLAD feature can be applied to image retrieval.

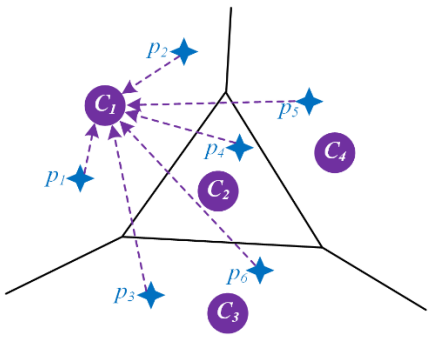

Fig. (a)

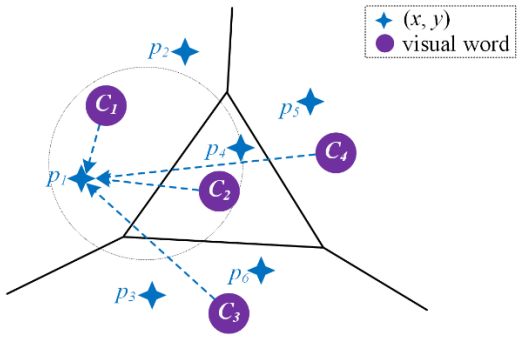

Fig. (b)

In order to clearly describe the VLAD module of this manuscript, we then illustrate the problem in the case of two-dimensional. As shown in Fig. (b), we take 6 2-dimensional points $\{p_i | i \in [1,6]\}$ as input for the VLAD module. Meanwhile, 4 visual words (cluster centers) are initialized, which are learnable parameters via backpropagation, denoted as $\{C_k | k \in [1,4]\}$. Each point p_i is assigned to each visual word C_k and represented by a residual vector $p_i - C_k$ recording the difference

between the point p_i and the visual word C_k. The relationship of the i-th point p_i to 4 visual words is denoted as $\boldsymbol{r}$. The (i, d) element of $\boldsymbol{r}$ is computed as follows:

$$\boldsymbol{r}(i,d) = \sum_{k=1}^{4} a_i(c_k)(p_i(d) - c_k(d)) \qquad d \in [1,2]$$

where $a_i(c_k)$ are the attention coefficients, $c_k(d)$ and $p_i(d)$ are the d-th dimension of the k-th visual word and the i-th point, respectively. The attention coefficients $a_i(c_k)$ are utilized to weight the importance of the i-th point to the k-th visual word.

Since each point just has strong relationships with a few visual words, it is necessary to simply consider the influence of visual words with higher attention scores on the high-level semantic feature. Therefore, we provide a top-k selection definition in Section 5.2.1. As shown in the Figure above, it is obvious that c_1 and c_2 have higher attention coefficients to the center point p_1 than that of c_3 and c_4. According to definition 5.1, we design a top-k VLAD feature selection operation (Note that we set top-k=2 in this case) for the VLAD module. Finally, the (i, d) element of $\boldsymbol{r}$ is computed as follows:

$$\boldsymbol{r}(i,d) = \sum_{k=1}^{2} a_i(c_k)(p_i(d) - c_k(d)) \qquad d \in [1,2]$$

Top-k controls the number of residual vectors on the one hand, and realizes overlap between different visual words on the other hand. At the same time, in order to increase the nonlinear transformation of the network, we use the shared FC layer, and finally aggregate the top-k transformed features, so the output of the VLAD module is consistent with the input (the dimensions of the vector are different). This not only accelerates the calculation but also improves the accuracy of the model.

[1] M. Angelina Uy and G. Hee Lee, Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4470–4479.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla and J. Sivic, NetVLAD: CNN architecture for weakly supervised place recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.