Notation: f_θ encoder (eg. ResNet50, DINOv2); g_θ aggregator (eg. VLAD, SALAD). Scalars may be lower or uppercase but are not bold. Matrices are always bold.

1 Input Image "Ablation"

Key idea: visualize how each part of the input image changes the output

$$\exists \mathbf{I} \in \mathbb{R}^{H \times W \times 3}, s \in \mathbb{R} \qquad P := \left\lceil \frac{H}{s} \right\rceil Q := \left\lceil \frac{W}{s} \right\rceil \tag{1}$$

$$\mathbf{M} := \{\mathbf{M}^{(pq)} := \begin{cases} 0, & (p-1)s \leq h \leq ps \wedge (q-1)s \leq w \leq qs \\ \mathbf{I}_{hw}, & otherwise \end{cases} \forall h \in [1,H], w \in [1,W] \mid \forall p \in [1,P], q \in [1,Q]\} \tag{2}$$

$$\mathbf{H} \in \mathbb{R}^{P \times Q},\ H_{pq} = \|(g_\theta \circ f_\theta)(\mathbf{M}^{(pq)}) - (g_\theta \circ f_\theta)(\mathbf{I})\| \tag{3}$$

Applicable to any VPR method in theory; however, fast inferencing speed is needed. Practical for *MixVPR*.

1. Divide $\mathbf{I}$ into square patches with side length s
2. Create pq masked images by setting pixels in each patch to 0
3. Difference between descriptors of image masked on patch pq and $\mathbf{I}$ is the activation at patch pq

2 Summation of Feature Importance

Key idea: sum the model's distribution of each feature to every cluster

$$\exists \mathbf{I} \in \mathbb{R}^{H \times W \times 3},\ \mathbf{F} := f_\theta(\mathbf{I}) \in \mathbb{R}^{C \times H' \times W'},\ \mathbf{S} := g'_\theta(\mathbf{I}) \in \mathbb{R}^{N \times H' \times W'} \tag{1}$$

$$\mathbf{F'} \in \mathbb{R}^{C \times (H' \cdot W')},\ \mathbf{F'}_c = ((col_1 \mathbf{F}_c^T)^T,\ \dots, (col_{H'} \mathbf{F}_c^T)^T) \in \mathbb{R}^{H' \cdot W'} \quad Similarly,\ \mathbf{S'} \in \mathbb{R}^{N \times (H' \cdot W')} \tag{2}$$

$$\mathbf{H'} \in \mathbb{R}^{C \times (H' \cdot W')}, \forall c \in [1,C]\ \mathbf{H'}_c = \sum_{i=1}^{N} \mathbf{F'}_c \odot \mathbf{S'}_i \tag{3}$$

$$\vec{h} = [\|col_1 \mathbf{H'}\|_2,\ \dots, \|col_C \mathbf{H'}\|_2] \quad or \quad \vec{h} = [\|col_1 \mathbf{S'}\|_2,\ \dots, \|col_C \mathbf{S'}\|_2] \tag{4}$$

$$\mathbf{H} \in \mathbb{R}^{H' \times W'} = [\vec{h}_{0:w}^T, \vec{h}_{w:2w}^T,\ \dots, \vec{h}_{(h-1)w:hw}^T] \tag{5}$$

1. Use a modified g_θ (g'_θ) that outputs a cluster assignment score $\mathbf{S}$ with N clusters
2. Flatten the last two dimensions of $\mathbf{F}$ and $\mathbf{S}$ in row-major order to get $\mathbf{F'}$ and $\mathbf{S'}$
3. **Anbang's algorithm:** for each channel of $\mathbf{F'}$' C channels, element-wise multiply by the feature-to-cluster score of every cluster in $\mathbf{S'}$' N clusters
4. *NetVLAD:* flattend activations $\vec{h}$ from channel dimension norm of $\mathbf{H'}$; *DINOv2SALAD:* use $\mathbf{S'}$ instead
5. Unflatten $\vec{h}$ to get the final activations matrix $\mathbf{H}$

3 Channel Sum of Feature Map

Key idea: straightforward adding across channel dimension

$$\exists \mathbf{I} \in \mathbb{R}^{H \times W \times 3},\ \mathbf{F} := f_\theta(\mathbf{I}) \in \mathbb{R}^{C \times H' \times W'} \qquad \mathbf{H} \in \mathbb{R}^{H' \times W'} = \sum_{i=1}^{C} \mathbf{F}_i$$

CricaVPR & SelaVPR can use this simple strategy as specified by their authors.