FAemb: a function approximation-based embedding method for image retrieval

Thanh-Toan Do, Quang D. Tran, Ngai-Man Cheung
Singapore University of Technology and Design (SUTD)

The objective of this paper is to design an effective embedding method mapping local features describing image (e.g. SIFT) to a higher dimensional representation used for image retrieval problem.

There is a wide range of methods [1, 2, 4, 5, 6, 7, 8] for finding a single vector to represent a set of local vectors proposed in the literature. Among these methods, VLAD [4] is a well-known embedding method used in image retrieval problem while TLCC [7] is one of successful embedding methods used in image classification problem.

VLAD and TLCC come from different motivations. VLAD's motivation is to characterize the distribution of residual vectors over Voronoi cells learned by a quantizer while TLCC's motivation is to *linearly approximate* a nonlinear function in high dimensional space, i.e., the nonlinear function $f(\mathbf{x})$ defined on $\mathbb{R}^d$ is approximated by $\mathbf{w}^T\phi(\mathbf{x})$ defined on $\mathbb{R}^D$ where $D > d$. Despite above differences, we show that VLAD is actually simplified version of TLCC. This means that we can depart from the idea of linear approximation of function to develop good embedding methods for image retrieval problem.

In TLCC, f is approximated using only its first order derivative information. In this paper, we propose to approximate f using higher order derivative information.

TLCC relied on the idea of coordinate coding defined bellow.

Definition 0.1 *Coordinate Coding [8]*
A coordinate coding of a point $\mathbf{x} \in \mathbb{R}^d$ *is a pair* $(\gamma(\mathbf{x}), \mathbf{C})$*, where* $\mathbf{C} = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{d\times n}$ *is a set of n anchor points, and* γ *is a map of* $\mathbf{x} \in \mathbb{R}^d$ *to* $\gamma(\mathbf{x}) = [\gamma_{\mathbf{v}_1}(\mathbf{x}), \ldots, \gamma_{\mathbf{v}_n}(\mathbf{x})]^T \in \mathbb{R}^n$ *such that* $\sum_{j=1}^n \gamma_{\mathbf{v}_j}(\mathbf{x}) = 1$*. It induces the following physical approximation of* $\mathbf{x}$ *in* $\mathbb{R}^d$*:* $\mathbf{x}' = \sum_{j=1}^n \gamma_{\mathbf{v}_j}(\mathbf{x})\mathbf{v}_j$*. A good coordinate coding should ensure that* $\mathbf{x}'$ *closes to* $\mathbf{x}$*.*

Our *F*unction *A*pproximation-based *emb*edding (FAemb) method is based on the following lemma.

Lemma 0.2 *If* $f: \mathbb{R}^d \to \mathbb{R}$ *is of class of* C^{k+1} *on* $\mathbb{R}^d$ *and* $\nabla^k f(\mathbf{x})$ *is Lipschitz continuous with constant* $M > 0$ *and* $(\gamma(\mathbf{x}), \mathbf{C})$ *is coordinate coding of* $\mathbf{x}$*, then*

$$\left| f(\mathbf{x}) - \sum_{j=1}^n \gamma_{\mathbf{v}_j}(\mathbf{x}) \sum_{|\alpha|\le k} \frac{\partial^\alpha f(\mathbf{v}_j)}{\alpha!}(\mathbf{x}-\mathbf{v}_j)^\alpha \right| \le \frac{M}{(k+1)!}\sum_{j=1}^n |\gamma_{\mathbf{v}_j}(\mathbf{x})| \left\|\mathbf{x}-\mathbf{v}_j\right\|_1^{k+1} \quad (1)$$

where α is multi-index notation [3].

If $k = 2$, then (1) becomes

$$\left| f(\mathbf{x}) - \sum_{j=1}^n \gamma_{\mathbf{v}_j}(\mathbf{x}) \left(f(\mathbf{v}_j) + \nabla f(\mathbf{v}_j)^T (\mathbf{x}-\mathbf{v}_j) + \frac{1}{2}\left(V\left(\nabla^2 f(\mathbf{v}_j)\right)\right)^T V\left((\mathbf{x}-\mathbf{v}_j)(\mathbf{x}-\mathbf{v}_j)^T\right)\right)\right| \le \frac{M}{6}\sum_{j=1}^n |\gamma_{\mathbf{v}_j}(\mathbf{x})| \left\|\mathbf{x}-\mathbf{v}_j\right\|_1^3 \quad (2)$$

where $V(\mathbf{A})$ is vectorization function flattening the matrix $\mathbf{A}$ to a vector by putting its consecutive columns into a column vector. ∇^2 is Hessian matrix.

The result derived from (2) is that the nonlinear function $f(\mathbf{x})$ can be approximated by linear form $\mathbf{w}^T\phi(\mathbf{x})$ where $\mathbf{w}$ can be defined as $\mathbf{w} = \left[\frac{1}{s_1}f(\mathbf{v}_j); \frac{1}{s_2}\nabla f(\mathbf{v}_j); \frac{1}{2}\left(V\left(\nabla^2 f(\mathbf{v}_j)\right)\right)\right]_{j=1}^n$ and the embedded vector $\phi(\mathbf{x})$-FAemb can be defined as

$$\phi(\mathbf{x}) = \left[s_1\gamma_{\mathbf{v}_j}(\mathbf{x}); s_2\gamma_{\mathbf{v}_j}(\mathbf{x})(\mathbf{x}-\mathbf{v}_j); \gamma_{\mathbf{v}_j}(\mathbf{x})V\left((\mathbf{x}-\mathbf{v}_j)(\mathbf{x}-\mathbf{v}_j)^T\right)\right]_{j=1}^n \in \mathbb{R}^{n(1+d+d^2)} \quad (3)$$

where s_1, s_2 are nonnegative scaling factors to balance three types of codes.

In order to get a good approximation of f, the RHS of (2) should be small enough. Furthermore, from definition of coordinate coding 0.1, $(\gamma(\mathbf{x}), \mathbf{C})$ should ensure that the reconstruction error $\|\mathbf{x}'-\mathbf{x}\|_2$ should be small. Putting two above criteria together, we find $(\gamma(\mathbf{x}), \mathbf{C})$ which minimize the following constrained objective function

$$Q(\gamma(\mathbf{x}), \mathbf{C}) = \|\mathbf{x}-\mathbf{C}\gamma(\mathbf{x})\|_2^2 + \mu\sum_{j=1}^n |\gamma_{\mathbf{v}_j}(\mathbf{x})| \left\|\mathbf{x}-\mathbf{v}_j\right\|_1^3 \quad st.\ \mathbf{1}^T\gamma(\mathbf{x}) = 1 \quad (4)$$

After learning $\mathbf{C}$ using training descriptors (e.g. minimizing (4) over training set), given a new descriptor $\mathbf{x}$, we get $\gamma(\mathbf{x})$ by minimizing (4) using learned $\mathbf{C}$. From $\gamma(\mathbf{x})$, we get the embedded vector $\phi(\mathbf{x})$-FAemb by using (3).

Table 1 presents results of our image retrieval framework using FAemb embedding method and the state of the art on Holidays and Oxford5k datasets. FAemb compares favorably with state-of-the-art embedding methods for image retrieval, such as VLAD, Fisher kernel, Temb, even with a shorter presentation.

Table 1: Comparison with the state of the art on Holidays and Oxford5k datasets. The frameworks are named by embedding methods used. n is number of anchor points. D is dimension of embedded vectors. Reference results are obtained from corresponding papers.

Framework	n	D	mAP Hol.	mAP Ox5k
VLAD [4]	256	16,384	58.7	-
Fisher [4]	256	16,384	62.5	-
VLAD_{LCS} [2]	64	8,192	65.8	51.7
$\text{VLAD}_{\text{intra}}$ [1]	64	8,192	56.5	44.8
$\text{VLAD}_{\text{intra}}$ [1]	256	32,536	65.3	55.8
$\text{VLAT}_{\text{improved}}$ [6]	64	9,000	70.0	-
Temb [5]	64	8,064	72.2	61.2
Temb [5]	128	16,256	73.8	62.7
Our framework				
FAemb	8	7,245	72.7	63.6
FAemb	16	15,525	**75.8**	**67.7**

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[1] Relja Arandjelovic and Andrew Zisserman. All about VLAD. In *CVPR*, 2013.
[2] Jonathan Delhumeau, Philippe Henri Gosselin, Hervé Jégou, and Patrick Pérez. Revisiting the VLAD image representation. In *MM*, 2013.
[3] Gerald B. Folland. *Advanced Calculus*. Prentice Hall, 1st edition, 2002.
[4] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local images descriptors into compact codes. *PAMI*, 2012.
[5] Hervé Jégou and Andrew Zisserman. Triangulation embedding and democratic aggregation for image search. In *CVPR*, 2014.
[6] Romain Negrel, David Picard, and Philippe Henri Gosselin. Web-scale image retrieval using compact tensor aggregation of visual descriptors. *IEEE Transactions on Multimedia*, 2013.
[7] Kai Yu and Tong Zhang. Improved local coordinate coding using local tangents. In *ICML*, 2010.
[8] Kai Yu, Tong Zhang, and Yihong Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.