Exemplar SVMs as Visual Feature Encoders

Joaquin Zepeda, Patrick Pérez
Technicolor

In this work, we investigate the use of *exemplar SVMs* (linear SVMs trained with one positive example only and a vast collection of negative examples) as encoders that turn generic image features into new, task-tailored features. The proposed feature encoding leverages the ability of the exemplar-SVM (E-SVM) classifier to extract, from the original representation of the exemplar image, what is unique about it. While existing image description pipelines rely on the intuition of the designer to encode uniqueness into the feature encoding process, our proposed approach does it explicitly relative to a "universe" of features represented by the generic negatives. We show that such a post-processing enhances the performance of state-of-the art image retrieval methods based on aggregated image features, as well as the performance of nearest class mean and K-nearest neighbor image classification methods. We establish these advantages for several features, including "traditional" features as well as features derived from deep convolutional neural nets. As an additional contribution, we also propose a recursive extension of this E-SVM encoding scheme (RE-SVM) that provides further performance gains.

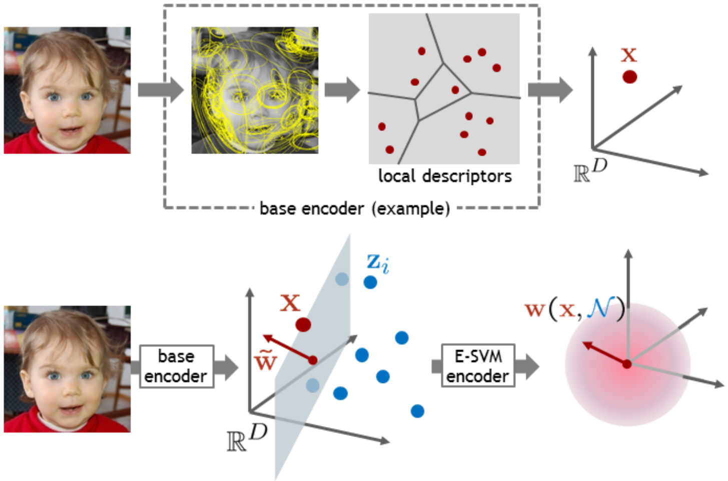

Figure 1: **Principle of Examplar SVM visual feature encoder.** (Top) Given a generic visual encoder, like BoW, Fisher vector or VLAD, an image is described as a fixed size feature vector $\mathbf{x} \in \mathbb{R}^D$; (Bottom) Using a pool of generic negative image features $\mathcal{N} = \{\mathbf{z}_i\}_{i=1}^N$, an E-SVM $\tilde{\mathbf{w}}$ is learned for each input image. The ℓ_2-normalized E-SVM $\mathbf{w}$ is the new encoding of the image for subsequent analysis.

Feature encoding with E-SVMs We assume that a generic, D-dimensional image feature encoder is given. This base encoder can be global, based on aggregated local features, or derived from CNNs-based features (Fig.1, top). We shall denote by vectors in $\mathbb{R}^D$ such features. An exemplar SVM can be computed from the exemplar feature vector $\mathbf{x}$ and a large set of generic feature vectors $\mathcal{N} = \{\mathbf{z}_i\}_{i=1}^N$ by solving the following optimization problem:

$$\tilde{\mathbf{w}}(\mathbf{x},\mathcal{N}) = \underset{\mathbf{w}\in\mathbb{R}^D}{\operatorname{argmin}}[\frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \alpha_+ \max(0, 1-\mathbf{x}^\top\mathbf{w}) + \alpha_- \sum_{i=1}^N \max(0, 1+\mathbf{z}_i^\top\mathbf{w})] \tag{1}$$

where λ, α_+ and α_- are positive parameters that control the level of regularization and the relative weight of negative examples. We will we refer to E-SVMs as the ℓ_2-normalized version of the solution to the above problem (Fig.1, bottom):

$$\mathbf{w}(\mathbf{x},\mathcal{N}) = \frac{\tilde{\mathbf{w}}(\mathbf{x},\mathcal{N})}{\|\tilde{\mathbf{w}}(\mathbf{x},\mathcal{N})\|_2}. \tag{2}$$

When dependence on $\mathbf{x}$ and $\mathcal{N}$ is clear from the context, we shall simply denote $\mathbf{w}$ this E-SVM.

Optimization problem (1) is a classic linear SVM problem relying on hinge loss, with the notable particularity that positive and negative sets are extremely unbalanced, one positive for up to, say, one million negatives. In [1], the property of hinge loss to yield dual solutions dependent only on a small number of (negative) support vectors is leveraged through hard negative mining. As an alternative efficient solver, we shall rely on stochastic gradient descent.

We propose using E-SVMs thus computed as new features. Hence we assume that we are given a first feature encoder, task-dependent or not, that produces feature vector $\mathbf{x}$ from a given image, but we instead use $\mathbf{w}(\mathbf{x},\mathcal{N})$ as the task-dependent feature representation for said image. Note that: (1) While E-SVM is a linear SVM, the resulting encoding is not linear relative to base feature $\mathbf{x}$; (2) This is a dimension preserving encoding, since the new image representation still lives in $\mathbb{R}^D$, in contrast with high-dimensional encoding (*e.g.*, using Fisher vectors [2] or explicit feature maps [3]).

Symmetric encoding for image search As demonstrated in [1], the E-SVM $\mathbf{w}^\circ = \mathbf{w}(\mathbf{x}^\circ,\mathcal{N})$ attached to a given image $\mathbf{x}^\circ$ can be used on its own to retrieve images with very similar content in a dataset $\mathcal{D} = \{\mathbf{x}_j\}_{j=1}^M$, using scores $\mathbf{x}_j^\top\mathbf{w}^\circ$. We propose instead a symmetric approach where each image $\mathbf{x}_j$ in the dataset is also equipped with its E-SVM feature $\mathbf{w}_j = \mathbf{w}(\mathbf{x}_j,\mathcal{N})$. Our approach then consists in sorting all these according to their similarity $s_j = \mathbf{w}_j^\top\mathbf{w}^\circ$ with the E-SVM of the query image.

Recursive E-SVMs encoding The above proposition of post-processing the output $\mathbf{x}$ of any generic feature encoder to produce E-SVM features $\mathbf{w}(\mathbf{x},\mathcal{N})$ suggests applying this procedure recursively. We can formalize this approach by first defining $\mathbf{w}^0 \triangleq \mathbf{x}$ and $\mathcal{N}^0 \triangleq \mathcal{N}$. The k-th recursion of E-SVM feature computation can then be written as follows for $k \geq 1$:

$$\mathbf{w}^k = \mathbf{w}(\mathbf{w}^{k-1},\mathcal{N}^{k-1}), \tag{3}$$

$$\text{where } \mathcal{N}^k = \{\mathbf{w}(\mathbf{z},\mathcal{N}^{k-1}), \mathbf{z} \in \mathcal{N}^{k-1}\}. \tag{4}$$

Features built using the k-th recursive E-SVM (RE-SVM-k) procedure specified in (3) can be used in a manner analogous to E-SMV to carry out image retrieval.

Performance gain For image search, a single RE-SVM recursion gives a large boost to performance obtained with VLAD-64, BoW-1000, Fisher-64 and CNN encodings (See mAP performance in Table below for VLAD-64, for instance), and a second iteration of E-SVM encoding yield additional gain.

	Holidays	Oxford 5K
VLAD-64	72.7	46.3
VLAD-64 + RE-SVM-1	77.5	55.5
VLAD-64 + RE-SVM-2	**78.3**	**57.5**

[1] T. Malisiewicz, A. Shrivastava, A. Gupta, and A. Efros. Exemplar-SVMs for visual object detection, label transfer and image retrieval. *ICML*, 2012.

[2] S. Jorge, F. Perronnin, and Z. Akata. Fisher vectors for fine-grained visual gategorization. *CVPR*, 2011.

[3] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE T-PAMI*, 34(3):480–92, 2012.

[4] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: generalizing to new classes at near-zero cost. *ECCV*, 2012.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. *NIPS*, 2012.

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