Maximum Margin Linear Classifiers in Unions of Subspaces

Xinrui Lyu[1,2]
xinrui.lyu@epfl.ch
Joaquin Zepeda[1]
joaquin.zepeda@technicolor.com
Patrick Pérez[1]
patrick.perez@technicolor.com

[1] Technicolor
35576, Cesson-Sevigne, France

[2] École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015, Lausanne, Switzerland

Dictionaries $\mathbf{D} \in \mathbb{R}^{d \times A}$ (with $A > d$) for sparse coding are learned in an unsupervised manner [4] by approximating the training vectors $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^M$ with sparse linear combinations of the columns (called *atoms*) of $\mathbf{D}$. Letting $\{\mathbf{z}_i \in \mathbb{R}^A\}_i$ denote the vectors of sparse linear combination weights, this can be formulated as

$$\underset{\mathbf{D}}{\operatorname{argmin}} \sum_{i=1}^{M} \min_{\mathbf{z}_i} \|\mathbf{x}_i - \mathbf{D}\mathbf{z}_i\|_2^2 + \beta\|\mathbf{z}_i\|_1. \quad (1)$$

We propose herein a novel method to learn SVM classifiers that picks up this line of work on dictionary learning by learning SVM classifiers that are sparse in a dictionary that is learned for the classification task. While previous works have addressed learning supervised dictionaries for classification, they have all focused on enforcing the sparsity of representation of the feature vectors and not of the classifiers, like we do.

Formulation. Given the dictionary $\mathbf{D}$, learning a linear classifier $\mathbf{w} = \mathbf{D}\mathbf{z}$ that is $\mathbf{D}$-sparse amounts to learning the sparse vector $\mathbf{z}$. A suitable formulation for $\mathbf{z}$ can be derived by substituting $\mathbf{w} = \mathbf{D}\mathbf{z}$ into the standard ℓ_2-penalized SVM learning objective and appending an additive sparsity-enforcing penalty term $\beta\|\mathbf{z}\|_1$. Our proposed dictionary learning problem follows by summing the resulting expression over K training classes:

$$\underset{\mathbf{D},\{\mathbf{z}_k\}_k,b}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{i=1}^{M} \max\left(0, 1 - y(\mathbf{x}_i^\top \mathbf{D}\mathbf{z}_k + b)\right) + \frac{\alpha}{2}\|\mathbf{D}\mathbf{z}_k\|_2^2 + \beta\|\mathbf{z}_k\|_1. \quad (2)$$

Note that $\mathbf{D}$ can further be fixed in latter stages and for never-before-seen clases where only the classifier's sparse $\mathbf{z}$ are learned. The learned linear classifiers $\mathbf{w} = \mathbf{D}\mathbf{z}$ will exist in a union of subspaces, with each subspace being the span of a small subset of atoms from $\mathbf{D}$. Hence we refer to our proposed classifier as a *Union-of-Subspaces SVM* (US-SVM). We further present two possible modifications of the above formulation. The first one forces the $\mathbf{z}_k$ to be non-negative, while the second one substitutes the ℓ_1 penalty term by an elastic net penalty term. The Stochastic Gradient Descent (SGD) solver we propose is valid for (2) and its variants incorporating one or both of the aforementioned modifications. It uses a block-coordinate descent approach for reasons of complexity, and adopts the ℓ_1-SVM SGD method described in [1].

Advantages of the proposed method. Forcing the classifier to be sparse in a learned dictionary exposes a number of interesting benefits. One benefit concerns the compactness of the representation – classifiers with compact representations can be stored more efficiently and, importantly, they incur lower computational cost both at training and testing time. Another benefit is that the atoms (columns) of the learned dictionary will inherit semantic properties shared by different classes and hence can often be interpreted as semantic *attributes*, thus opening a possible path to weakly supervised attribute discovery. In a similar manner, atoms of the learned dictionary will often correspond to modalities of the underlying feature distribution that can likewise have interesting semantic interpretations. Forcing the classifier to be sparse using a learned dictionary can also be interpreted as a novel SVM regularization scheme. Unlike other schemes that constrain the norm of the classifier, our regularization requires that all classifiers be represented in terms of a common dictionary, in effect enabling the system to leverage the annotations for all classes when learning any given class.

Experiments. We evaluate our proposed method using both unsupervised features as well as very recent, CNN-derived features, testing it on well known image classification datasets (PASCAL VOC 2007 [3] and ImageNet [2]). Our experiments establish that our approach results in very sparse representations of classifiers that outperform other SVM classifiers, and with learned dictionaries that carry out automatic attribute and modality discovery as part of the learning process. Example results on a subset of ImageNet are presented in Fig. 2 for US-SVM, with/without non-negativity constraints and elastic net penalization.

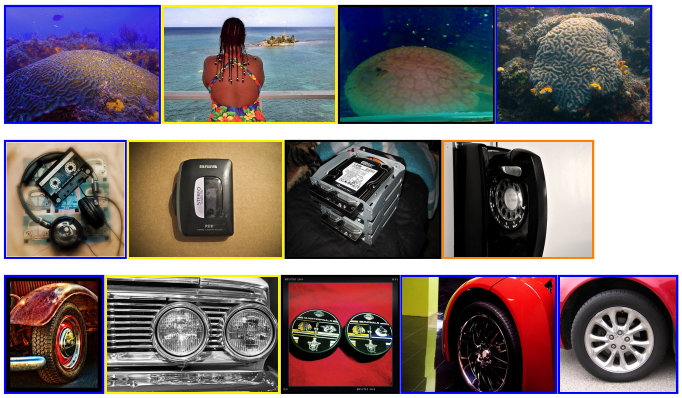

Figure 1: Examples of visual attributes captured by learned atoms across multiple classes. *Top*: water; *middle*: rectangular shape; *bottom*: round shape. For each row, images with the same border color belong to the same class.

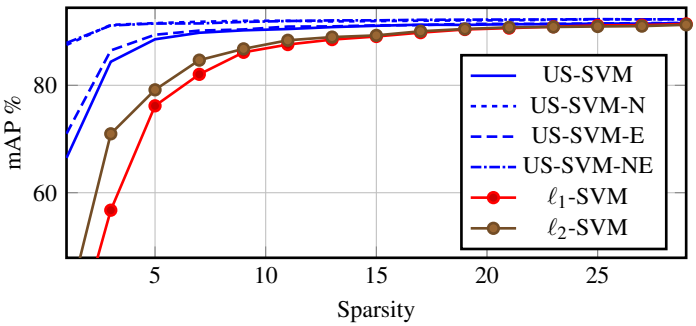

Figure 2: Classification performance versus sparsity (average $|\mathbf{z}_k|_0$) on a subset of ImageNet for US-SVMs with and without Non-negativity (N) constraints and Elastic net (E) penalization, and ℓ_1/ℓ_2-SVMs.

In Fig. 2 we present example results on ImageNet that illustrate how US-SVM enjoys nearly constant performance for drastically low sparsity levels of < 5 (for feature vectors $\mathbf{x}_i \in \mathbb{R}^{128}$), where the performance is close to 20 mAP points better (a $+20\%$ difference) than that of ℓ_1-SVM or ℓ_2-SVM.

Attribute discovery. Fig. 1 shows examples of visual "attributes" (associated to atoms) that are learned automatically using our proposed method. Each row of images corresponds to the top ranked images when using the corresponding atom alone as a classifier. Note that the classes of these top-ranked images are very varied, indicating that the atom has not specialized to a given class. Yet common visual elements are evident for the top-ranked images of each row, whether geometric (round or rectangular parts) or texture-based (water). Although not illustrated, we further show empirically that some learned atoms further split image classes into the various modalities of the class.

[1] Leon Bottou. Stochastic gradient descent tricks. In Grégoire Montavon, Geneviève Orr, and Klaus-Rober Müller, editors, *Neural Networks: Tricks of the Trade*, volume 1. Springer, 2 edition, 2012.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009.

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[4] Joaquin Zepeda, Christine Guillemot, and Ewa Kijak. Image Compression Using Sparse Representations and the Iteration-Tuned and Aligned Dictionary. *IEEE Transactions on Signal Processing*, 5(5):1061–1073, 2011.