Combining complementary kernels in complex visual categorization

Nicolas Thome[1], David Picard[2] and Matthieu Cord[1]
[1]LIP6 UPMC Paris 6
4 place Jussieu, 75005 Paris, France
[2]ETIS, CNRS ENSEA UCP, 95000 Cergy-Pontoise

1 Introduction

Bag Of Words model [1] and Fisher Vectors [2] coupled with incorporated spatial information, such as the Spatial Pyramid Matching (SPM) [3] or the spatial Fisher vectors [4], proved to reach state of the art performances in many Image categorization tasks, *e.g.* the PASCAL VOC challenge [5]. However, image categorization remains a very challenging task because most descriptors present strong intra-class variabilities and inter-class correlations. Therefore, a natural way to improve categorization performances consists in designing efficient feature combination strategies. It is an important issue for both computer vision and machine learning communities, that has been extensively studied in the last decade. Multiple Kernel Learning (MKL) is appealing for that purpose, since it offers the possibility to jointly learn the weighting of the different channels (features and similarity kernels) and the classification function [6]. The goal is to find the optimal classification function f defined as follows: $f(\mathbf{x}) = \sum_i \alpha_i y_i \sum_m \beta_m k_m(\mathbf{x}, \mathbf{x}_i) - b$ where the variable to be optimized are both the α and the w. Efficient algorithms exist for solving the related optimization convex problem [7].

Recent works attempting at using MKL on image datasets for combining different channels [8, 9] use MKL optimization algorithms based on ℓ_1 norm to regularize the kernel weights, like SimpleMKL [7]. Since this leads to sparse solutions, most studies report that MKL is often outperformed by simple baseline methods (product or averaging) [8, 9]. However, especially in our Computer Vision context, the different kernels are generated from different visual modalities, most of them being informative and many of them being complementary (*e.g.* edge, color and texture). Therefore, we are not interested in performing kernel selection, but we aim at finding a proper weighting between them. There exists however ℓ_2 MKL optimization schemes [10] to solve the MKL problem, but except [11], there has been few attempt to apply these schemes on image databases to find a non-sparse combination of complementary descriptors.

2 Propositions

For this workshop, we would like to propose two main algorithms, several experiments and feedback discussion about successes and failures of the different strategies we have investigated. The two kernel-learning proposed algorithms are:

1. An hybrid strategy already published in [12] that attempts at learning a non-sparse combination between different image modalities, but still using a ℓ_1 optimization algorithm. The idea is to generate for each descriptor numerous kernels by varying their parameters (*e.g.* standard deviation σ for gaussian kernels). Thus, for each channel c, we form a set of M kernels $K_{c,\sigma}$, and use a ℓ_1 MKL strategy to select the relevant σ parameter. Our adapted MKL problem formulation leads to find the optimal function of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{N_e} \alpha_i y_i \sum_{c=1}^{N_c} \sum_{\sigma=\sigma_1}^{\sigma_M} \beta_{c,\sigma} k_{c,\sigma}(\mathbf{x}, \mathbf{x}_i) - b \quad (1)$$

   where the joint optimization is performed on α_i (N_e parameters) and $\beta_{c,\sigma}$ ($N_c \times M$ parameters).

   One interesting feature of the approach is the ability to jointly learn individual kernel parameters σ and kernel combination coefficients β_m. The sparse solution output by ℓ_1 MKL algorithms is therefore used as an option to cross-validation. Other approaches like ℓ_2 MKL use a two-step procedure: optimal σ is first determined by cross-validation, and combining the kernels is then performed for a fixed σ. This leads to a sub-optimal parameter estimation with respect to our global optimization scheme.

2. An unpublished work to learn a powered product of kernels, denoted **Product Kernel Learning** (PKL). In the case of redundant features, the product of associated kernels is shown to achieve good results [9]. PKL

is a further refinement by considering a geometric combination of kernels: $K(\mathbf{x}_1, \mathbf{x}_2) = \prod_c k_c(\mathbf{x}_1, \mathbf{x}_2)^{\beta_c}$. The classification function is thus: $f(x) = \sum_i \alpha_i y_i \prod_c k_c(\mathbf{x}_i, \mathbf{x})^{\beta_c} - b$.

Like in MKL we aim at jointly learning α_i and β_c parameters. We restrain in our experiments the combination to exponential kernels: $k_c(\mathbf{x}_1, \mathbf{x}_2) = e^{-\gamma d_c(\mathbf{x}_1, \mathbf{x}_2)}$, with d being the distance relative to the c^{st} feature, so that the classification function may be written as: $f(\mathbf{x}) = \sum_i \alpha_i y_i \prod_c e^{-\gamma_c d_c(\mathbf{x}_1, \mathbf{x}_2)}$. We propose an efficient algorithm to jointly learn the parameters α_i and γ_c, using an alternate optimization scheme: the α_i are learned using a standard SMO solver (on the dual problem), while the γ_c are optimized with an approximate second order gradient descent. The convexity of this optimization problem will be discussed too.

3 Experiments

We provide evaluations of our algorithms on the following datasets:

- Sonar and Ionosphere UCI, standard setups in the machine learning community. In these databases, PKL outperforms MKL [7] with a large margin.
- Computer Vision databases: Caltech, Oxford Flowers, and PASCAL VOC2009. We evaluate the performances of feature combination using edge, color and texture descriptors. We compare different combination methods: baseline methods, (averaging & product), weighted sum kernel, our adapted MKL approach, and our PKL method. We comment successes and failures obtained for different categorization tasks using baseline and advanced kernel learning methods for the different databases. In particular, we qualify the conclusions provided by recent studies [8, 9] about combining complementary descriptors with large performance variations.

Additionally, we report experiments using PKL formalism to learn visual words in a discriminative way. We comment our results that are contrasted with respect to [13]. We will also discuss on how quantifying the feature redundancy in order to evaluate its impact for combination performances. This aspect may be decisive to guide the choice of the sum vs product combination of kernels.

References

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, Oct. 2003, vol. 2, pp. 1470–1477.

[2] Florent Perronnin and Christopher R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.

[3] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR '06*, Washington, DC, USA, 2006, pp. 2169–2178, IEEE Computer Society.

[4] J. Krapac, J. Verbeer, and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *ICCV*, 2011.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results," http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[6] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *ICML '04*, 2004, p. 6.

[7] Alain Rakotomamonjy, Francis Bach, Stephane Canu, and Yves Grandvalet, "SimpleMKL," *JMLR*, vol. 9, pp. 2491–2521, 2008.

[8] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *ICCV*, 2009.

[9] Peter V. Gehler and Sebastian Nowozin, "On feature combination for multiclass object classification," in *IEEE ICCV*, 2009.

[10] Marius Kloft, Ulf Brefeld, Soeren Sonnenburg, Pavel Laskov, Klaus-Robert Müller, and Alexander Zien, "Efficient and accurate lp-norm multiple kernel learning," in *NIPS*, 2009, pp. 997–1005.

[11] Fei Yan, Krystian Mikolajczyk, Josef Kittler, and Muhammad Tahir, "A comparison of l1 norm and l2 norm multiple kernel svms in image and video classification," *CBMI, International Workshop on*, vol. 0, pp. 7–12, 2009.

[12] David Picard, Nicolas Thome, and Matthieu Cord, "An efficient system for combining complementary kernels in complex visual categorization tasks," in *ICIP*, 2010, pp. 3877–3880.

[13] H. Cai, F. Yan, and K. Mikolajczyk, "Learning weights for codebook in image classification and retrieval," in *CVPR*, 2010.