Scene Classification with Semantic Fisher Vectors

Mandar Dixit[1], Si Chen[1], Dashan Gao[2] Nikhil Rasiwasia[3] Nuno Vasconcelos[1]
[1]University of California, San Diego. [2]Qualcomm Inc., San Diego. [3]SnapDeal.com, India.

Semantic image classification has been a topic of significant interest in computer vision. Many authors have argued for the merit of semantically meaningful features instead of low level filter responses or mid level codes for vision tasks [1, 3, 4, 5, 8]. For scene classification, in particular, images are often represented as *Bags of semantics* (BoS). This is achieved by scoring image regions with the help of trained concept labelers (ex. object detectors) [3, 5, 8]. Despite the potential benefits of the BoS, it has not been very successful so far, for two main reasons: 1) The absence of strong labelers producing reliable semantics. 2) A lack of an invariant embedding (ex. Fisher vectors [6, 7]) specifically suited for semantic descriptors. The recent advance in object classification using CNNs [2] has solved the problem of noisy semantic labeling. The problem of deriving an invariant scene representation with CNN semantics, however, is a challenging task and our main focus in this paper.

The CNN in [2] which is pre-trained for a vocabulary $\mathcal{V} = \{v_1, \ldots, v_S\}$ of S *semantic concepts* (1000 ImageNET classes), generates the image BoS $I = \{\pi_1, \pi_2, \ldots, \pi_n\}$ producing, for each image patch, a posterior probability vector π_i. These vectors or *semantic descriptors* are multinomial parameters and their distribution in an image can be modeled using a Dirichlet mixture. This inspired our first proposal for an embedding of image BoS, the Dirichlet Mixture Fisher Vector (DMM-FV). The log likelihood of an image BoS under a DMM can be expressed as,

$$\mathcal{L} = \log P(\{\pi_i\}_{i=1}^{n} | \{\alpha_k, w_k\}_{k=1}^{K}) \quad (1)$$

$$= \log \prod_{i=1}^{n} \sum_{k=1}^{K} w_k \frac{\gamma(\sum_l \alpha_{kl})}{\prod_l \gamma(\alpha_{kl})} e^{\sum_l (\alpha_{kl} - 1) \log \pi_{il}}. \quad (2)$$

where α_k, w_k are the distribution parameters, and $\gamma(x) = \int_0^\infty x^{t-1} e^{-x} dx$. The Fisher scores are easily obtained as,

$$\mathcal{G}^I_{\alpha_k} = \frac{1}{n} \sum_{i=1}^{N} p(k|\pi_i) \left(\psi(\sum_l \alpha_{kl}) - \psi(\alpha_k) + \log \pi_i \right) \quad (3)$$

where $\psi(x) = \frac{\partial \gamma(x)}{\partial x}$. Using (3) and an appropriate Fisher information matrix, the image BoS can be encoded into a DMM FV.

Since, Dirichlet distribution is a natural model for probability vectors, it is natural to expect that DMM-FV would perform as impressively on semantics as the classic Gaussian Mixture Fisher vector (GMM-FV) does with SIFT [7]. Our experiments, however, reveal that the solution fails miserably most likely due to the complicated nature of the space of probability vectors (simplex). Classifiers generate probabilities by propagating features through a sigmoid or a soft-max function. These transformations are highly non-linear and therefore destroy the Euclidean properties of the original space. This is true for the simplex which is characterized by non-metrics like KL divergence and has non-linear geodesics. Therefore, mixture modeling on this non-Euclidean space is likely to be very difficult.

To circumvent the problems presented by the simplex geometry, we seek an alternative interpretation of semantic descriptors π's. Like the parameters of any exponential family, these multinomials can be expressed in their natural parameter form $\nu = \eta(\pi)$. When the semantics are binary, the natural parameter is obtained by a logit transform $\nu = \log \frac{\pi}{1-\pi}$. This is shown to map the features from a highly-nonlinear semantic space back into a simpler Euclidean space. For multinomials, this transformation takes the forms,

$$\nu_k^{(1)} = \log \pi_k \quad (4)$$

$$\nu_k^{(2)} = \log \pi_k + C \quad (5)$$

$$\nu_k^{(3)} = \log \frac{\pi_k}{\pi_S} \quad (6)$$

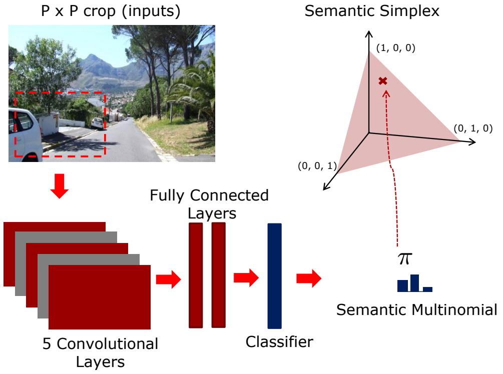

Figure 1: CNN based semantic image representation. Each image patch is mapped into an SMN π on the semantic space $\mathcal{S}$ by the ImageNET CNN. The resulting image representation is known as a Bag of semantics.

where ν_k and π_k are the k^{th} entries of ν and π, respectively. Since the natural parameters ν, unlike semantic multinomials π are likely to reside in a simpler space, even a standard Gaussian Mixture FV [6] can be used to a good effect.

With the help of the natural parameter formulations (4)- (6) for ImageNET semantics and a GMM FV, we derive our BoS based scene representation. We refer to it simply as a *semantic Fisher vector*. The semantic FV is shown to be better than FVs of intermediate layer CNN features due to its invariance, which is a direct result of semantic abstraction. We show that a classifier of semantic FV outperforms even a fine-tuned ImageNET CNN. The proposed semantic FV relies on object semantics. As an image representation, therefore, it is complementary to the features from the scene classification network (Places CNN) recently proposed in [9]. Our experiments show that a simple combination of the two descriptors, produces a state-of-the-art scene classifier for MIT Indoor and MIT SUN benchmarks.

[1] Alessandro Bergamo and Lorenzo Torresani. Classemes and other classifier-based features for efficient object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2014.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[3] Roland Kwitt, Nuno Vasconcelos, and Nikhil Rasiwasia. Scene recognition on the semantic manifold. In *Proceedings of the 12th European conference on Computer Vision - Volume Part IV*, ECCV'12, pages 359–372, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-33764-2. doi: 10.1007/978-3-642-33765-9_26.

[4] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.140.

[5] Li-Jia Li, Hao Su, Yongwhan Lim, and Fei-Fei Li. Object bank: An object-level image representation for high-level visual recognition. *International Journal of Computer Vision*, 107(1):20–39, 2014.

[6] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15560-X, 978-3-642-15560-4.

[7] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[8] Yu Su and Frédéric Jurie. Improving image classification using semantic attributes. *International Journal of Computer Vision*, 100(1):59–77, 2012.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.

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