Scene Classification with Semantic Fisher Vectors

Mandar Dixit[†], Si Chen[†], Dashan Gao[‡], Nikhil Rasiwasia[§] and Nuno Vasconcelos[†]
[†]University of California, San Diego
[‡]Qualcomm Inc., San Diego
[§]SnapDeal.com, India
{mdixit, sic046, nvasconcelos}@ucsd.edu, dgao@qti.qualcomm.com,
nikhil.rasiwasia@gmail.com

Abstract

With the help of a convolutional neural network (CNN) trained to recognize objects, a scene image is represented as a bag of semantics (BoS). This involves classifying image patches using the network and considering the class posterior probability vectors as locally extracted semantic descriptors. The image BoS is summarized using a Fisher vector (FV) embedding that exploits the properties of the space of these descriptors. The resulting representation is referred to as a semantic Fisher vector. Two implementations of a semantic FV are investigated. First involves modeling the BoS with a Dirichlet Mixture and computing the Fisher gradients for this model. Due to the difficulty of mixture modeling on a non-Euclidean probability simplex, this approach is shown to be unsuccessful. A second implementation is derived using the interpretation of semantic descriptors as parameters of a multinomial distribution. Like the parameters of any exponential family, these can be projected into their natural parameter space. For a CNN, this is shown equivalent to using inputs of its soft-max layer as patch descriptors. A semantic FV is then computed as a Gaussian Mixture FV in the space of these natural parameters. This representation is shown to outperform other alternatives such as FVs of features from the intermediate CNN layers or a classifier obtained by adapting (fine-tuning) the CNN. The proposed FV represents an embedding for object classification probabilities. As an image representation, therefore, it is complementary to the features obtained from a scene classification CNN. A combination of the two representations is shown to achieve state-of-the-art results on MIT Indoor scenes and SUN datasets.

1. Introduction

Natural scene classification is a challenging problem for computer vision, since most scenes are collections of entities (e.g. objects) organized in a highly variable layout. This high variability in appearance has made flexible visual representations quite popular for this problem. Many works have proposed to represent scene images as orderless collections, or "bags," of locally extracted visual features, such as SIFT or HoG [23, 5]. This is known as the bag-of-features (BoF) representation. For the purpose of classification, these features are pooled into an invariant image representation known as the Fisher vector (FV) [12, 25], which is then used for discriminant learning. Until very recently, bag-of-SIFT FV achieved state-of-the-art results for scene classification [30].

Recently, there has been much excitement about alternative image representations, learned with convolutional neural networks (CNNs) [19], which have demonstrated impressive results on large scale object recognition [16]. This has prompted many researchers to extend CNNs to problems such as action recognition [15], object localization [9], scene classification [10, 39] and domain adaptation [8]. Current multi-layer CNNs can be decomposed into a first stage of convolutional layers, a second fully-connected stage, and a final classification stage. The convolutional layers perform pixel wise transformations, followed by localized pooling, and can be thought of as extractors of visual features. Hence, the convolutional layer outputs are a BoF representation. The fully connected layers then map these features into a vector amenable to linear classification. This is the CNN analog of a Fisher vector mapping.

Beyond SIFT Fisher vectors and CNN layers, there exists a different class of image mappings known as *semantic representations*. These mappings require vectors of classifier outputs, or *semantic descriptors,* extracted from an image. Several authors have argued for the potential of such representations [35, 27, 33, 17, 18, 3, 20]. For example, semantic representations have been used to describe objects by their attributes [18], represent scenes as collections of objects [20] and capture contextual relations between classes [29]. For some visual tasks, such as hashing or large scale retrieval, a global semantic descriptor is usually preferred [34, 4]. Proposals for scene classification, on the other hand, tend to rely on a collection of locally extracted semantic image descriptors, which we refer to as bag of semantics (BoS) [33, 17, 20]. While a BoS based scene representation has outperformed low-dimensional BoF rep-

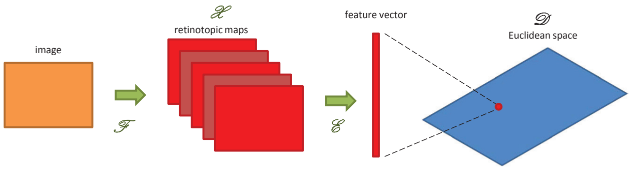

Figure 1. Bag of features (BoF). A preliminary feature mapping $\mathcal{F}$, maps an image into a space $\mathcal{X}$ of retinotopic features. A non-linear embedding $\mathcal{E}$ is then used to map this intermediate representation into a feature vector on an Euclidean space $\mathcal{D}$.

resentations [17], it is usually less effective than the high dimensional BoF-FV. This is due to the fact that, 1) local or patch-based semantic features can be very noisy, and 2) it is harder to combine them into a global image representation, akin to the Fisher vector.

In this work, we argue that highly accurate classifiers, such as the ImageNET trained CNN of [16] eliminate the first problem. We obtain a BoS image representation using this network by extracting semantic descriptors (object class posterior probability vectors) from local image patches. We then consider the design of a *semantic Fisher vector*, which is an extension of the standard Fisher vector to this BoS. We show that this is difficult to implement directly on the space of probability vectors, because of its non-Euclidean nature. On the other hand, if semantic descriptors from an image are seen as parameters of a multinomial distribution and subsequently mapped into their natural parameter space, a robust semantic FV can be obtained simply using the standard Gaussian mixture based encoding of the transformed descriptors [25]. In case of a CNN, this natural parameter mapping is shown equivalent to the inverse of its soft-max function. It follows that the semantic FV can be implemented as a classic (Gaussian Mixture) FV of pre-softmax CNN outputs.

The semantic FV, represents a strong embedding of features that are fairly *abstract* in nature. Due to the invariance of this representation, which is a direct result of semantic abstraction, it is shown to outperform Fisher vectors of lower layer CNN features [10] as well as a classifier obtained by fine-tuning the CNN itself [9]. Finally, since object semantics are used to produce our image representation, it is complementary to the features of the scene classification network (Places CNN) proposed in [39]. Experiments show that a simple combination of the two descriptors, produces a state-of-the-art scene classifier on MIT Indoor and MIT SUN benchmarks.

2. Image representations

In this section, we briefly review BoF and BoS based image classification.

2.1. Bag of Features

Both the SIFT-FV classifier and the CNN are special cases of the general architecture in Figure 1, commonly known as the bag of features (BoF) classifier. For an image $I(l)$, where l denotes spatial location, it defines an initial mapping $\mathcal{F}$ into a set of *retinotopic* feature maps $f_k(l)$. These maps preserve the spatial topology of the image. Common examples of mapping $\mathcal{F}$ include dense SIFT, HoG and the convolutional layers of a CNN. The BoF produced by $\mathcal{F}$ is subject to a highly nonlinear *embedding* $\mathcal{E}$ into a high dimensional feature space $\mathcal{D}$. This is a space with Euclidean structure, where a linear classifier $\mathcal{C}$ suffices for good performance.

It could be argued that this architecture is likely to always be needed for scene classification. The feature mapping $\mathcal{F}$ can be seen as a (potentially non-linear) local convolution of the input image with filters, such as edge detectors or object parts. This enables the classifier to be highly selective, e.g. distinguish pedestrians from cars. However, due to its retinotopic nature, the outputs of $\mathcal{F}$ are sensitive to variations in scene layout. The embedding $\mathcal{E}$ into the non-retinotopic space $\mathcal{D}$ is, therefore, necessary for *invariance* to such changes. Also, the space $\mathcal{D}$ must have a Euclidean structure to support classification with a linear decision boundary.

CNN based classifiers have recently achieved spectacular results on the ImageNET object recognition challenge [16, 31]. Their success has encouraged many researchers to use the features and embeddings learned by these networks for scene classification, replacing the traditional SIFT-FV based architecture [8, 32, 10, 22]. It appears undisputable that their retinotopic mapping $\mathcal{F}$, which is strongly non-linear (multiple iterations of pooling and rectification) and discriminant in nature (due to backpropagation) [37], has a degree of selectivity that cannot be matched by shallower mappings, such as SIFT. Less clear, however, is the advantage of using embeddings learned on ImageNET in place of the Fisher vectors for scene representation. As scene images exhibit a greater degree of intra class variation compared to object images, the ability to trade-off selectivity with invariance is critical for scene classification. While Fisher vectors derived using mixture based encoding are invariant by design, a CNN embedding learned from almost centered object images is unlikely to cope with the variability in scenes.

2.2. Bag of Semantics

Semantic representations are an alternative to the architecture of Figure 1. They simply map each image into a set of classifier outputs, using these as features for subsequent processing. The resulting feature space $\mathcal{S}$ is commonly known as the *semantic feature space*. Since scene semantics vary across image regions, scene classification requires a spatially localized semantic mapping. This is denoted as the *bag-of-semantics* (BoS) representation.

As illustrated in Figure 2, the BoS is akin to the BoF, but based on semantic descriptors. Its first step is the retinotopic maping $\mathcal{F}$. However, instead of the embedding $\mathcal{E}$, this is followed by another *retinotopic* mapping $\mathcal{N}$ into $\mathcal{S}$. At

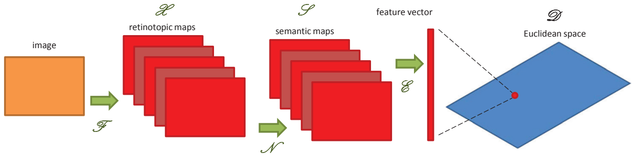

Figure 2. Bag of semantics (BoS). The space $\mathcal{X}$ of retinotopic features is first mapped into a retinotopic semantic space $\mathcal{S}$, using a classifier of image patches. A non-linear embedding $\mathcal{E}$ is then used to map this representation into a feature vector on an Euclidean space $\mathcal{D}$.

each location l, $\mathcal{N}$ maps the BoF descriptors extracted from a neighborhood of l into a *semantic descriptor*. The dimensions of this descriptor are probabilities of occurrence of visual classes (e.g. object classes, attributes, etc.). A BoS is an ensemble of retinotopic maps of these probabilities. An embedding $\mathcal{E}$ is used to finally map the BoS features into a Euclidean space $\mathcal{D}$.

While holistic semantic representations have been successful for applications like image retrieval or hashing, localized representations, such as the BoS, have proven less effective for scene classification, for a couple of reasons. First, the scene semantics are *hard to localize*. They vary from image patch to image patch and it has been difficult to build reliable scene patch classifiers. Hence, local semantics tend to be noisy [28, 20] and most works use a single global semantic descriptor per image [34, 2, 3]. This may be good for hashing, but it is not expressive enough for scene classification. Second, when semantics are extracted locally, the embedding $\mathcal{E}$ into an Euclidean space has been difficult to implement [17]. This is because semantic descriptors are probability vectors, and thus inhabit a very non-Euclidean space, the probability simplex, where commonly used descriptor statistics lose their effectiveness. In our results we show that even the sophisticated Fisher vector encoding [25], when directly implemented, has poor performance on this space.

We argue, that the recent availability of robust classifiers such as the CNN of [16], trained on large scale datasets, such as ImageNET [7], effectively solves the problem of noisy semantics. This is because an ImageNET CNN is, in fact, trained to classify objects which may occur in local regions or patches of a scene image. The problem of implementing an invariant embedding $\mathcal{E}$ in the semantic space, however, remains to be solved.

3. BoF embeddings

We first try to analyze, the suitability for scene classification, of the known BoF embeddings, namely the Fisher vector and the fully connected layers of ImageNET CNNs.

3.1. CNN embedding

For the CNN of [16], the mapping $\mathcal{F}$ consists of 5 convolutional layers. These produce an image BoF $\mathcal{I} = \{x_1, x_2, \ldots x_N\}$, where x_i's are referred to as the *conv5* descriptors. The descriptors are max pooled in their local neighborhood and transformed by the embedding $\mathcal{E}$. The embedding is implemented using two fully connected network stages, each performing a linear projection, and a non-linear *ReLu* transformation $\{W \times (.)\}_+$. The resulting outputs of layer 7, which we denote as *fc7*, are the features of space $\mathcal{D}$, in Figure 1.

3.2. FV embedding

Alternatively, a FV embedding can be implemented for the BoF of *conv5* descriptors. This consists of a preliminary projection into a principal component analysis (PCA) subspace,

$$x = Cz + \mu, \tag{1}$$

where C is a low-dimensional PCA basis and z are the coefficients of projection of the *conv 5* descriptors x on it. z's are assumed Gaussian mixture distributed.

$$z \sim \sum_k w_k N(\mu_k, \sigma_k). \tag{2}$$

A central component of the FV is the natural gradient with respect to parameters (mean, variance and weights) of this model [30]. For *conv5* features, we have found that the gradient with respect to the mean [25]

$$\mathcal{G}^{\mathcal{I}}_{\mu_k} = \frac{1}{N\sqrt{w_k}} \sum_{i=1}^{N} p(k|z_i) \left(\frac{z_i - \mu_k}{\sigma_k} \right) \tag{3}$$

suffices for good performance. Note that this gradient is an encoding and pooling operation over the z_i. It destroys the retinotopic topology of the BoF and guarantees invariance to variations of scene layout.

3.3. Comparison

We compared the CNN and FV embeddings, on two popular object recognition (Caltech 256 [11]) and scene classification (MIT Indoors [26]) datasets, with the results shown in the top half of Table 1. For the CNN embedding, 7^{th} fully connected layer features were obtained with "Caffe" [14]. Following [8], this 4096 dimensional feature vector was extracted globally from each image. It was subsequently power normalized (square rooted), and L_2 normalized, for better performance [32]. The classifier trained with this representation is denoted "fc 7" in the table. For the FV embedding, the 256-dimensional *conv5* descriptors were PCA reduced to 200 dimensions and pooled with (3), using a 100-Gaussian mixture. This was followed by a square root and L_2 normalization, plus a second PCA to reduce dimensionality to 4096 and is denoted "conv5 + FV" in the table. Both representations were used with a linear SVM classifier.

The results of this experiment highlight the strengths and limitations of the two embeddings. While *fc7* is vastly superior to the FV for object recognition (a gain of almost 12% on Caltech), it is clearly worse for scene classification (a loss of 2% on MIT Indoor). This suggests that, although

Table 1. Comparison of the ImageNET CNN and FV embeddings on scene and object classification tasks.

Method	MIT Indoor	Caltech 256
fc 7	59.5	68.26
conv5 + FV	61.43	56.37
fc7 + FV	65.1	60.97

invariant enough to represent images containing single objects, the CNN embedding cannot cope with the variability of the scene images. On the other hand, the mixture based encoding mechanism of the FV is quite effective on the scene dataset.

FV over *conv 5*, however, is an embedding of low-level CNN features. In principle, an equivalent embedding of BoS features should have better performance, since semantic descriptors have a higher level of abstraction than *conv5*, and thus exhibit greater invariance to changes in visual appearance. To some extent, the image representation proposed by Gong *et al*. [10] shows the benefits of such invariance, albeit using an embedding of the intermediate 7^{th} layer activations, not the semantic descriptors at the network output. They represent a scene image as a collection of *fc7* activations extracted from local crops or patches. These are summarized using an approximate form of (3), known as VLAD [13]. The resulting embedding, denoted as “fc7 + FV” in Table 1, is very effective for scene classification[1]. However, since the representation does not derive from semantic features, it is likely to be both less discriminant and less abstract than the truly semantic embedding of Figure 2. The implementation of an effective semantic embedding, on the other hand, is not trivial. We consider this problem in the remainder of this work.

4. Semantic FV embedding

We start with a brief review of a BoS image representation and then propose suitable embeddings for them.

4.1. The BoS

Given a vocabulary $\mathcal{V} = \{v_1, \ldots, v_S\}$ of S *semantic concepts*, an image I can be described as a bag of instances from these concepts, localized within image patches/regions. Defining an S-dimensional binary indicator vector s_i, such that $s_{ir} = 1$ and $s_{ik} = 0$, $k \neq r$, when the i^{th} image patch x_i depicts the semantic class r, the image can be represented as $I = \{s_1, s_2, \ldots, s_n\}$, where n is the total number of patches. Assuming that s_i is sampled from a multinomial distribution of parameter π_i, the log-likelihood of image I can be expressed as,

$$\mathcal{L} = \log \prod_{i=1}^{n} \prod_{r=1}^{S} \pi_{ir}{}^{s_{ir}} = \sum_{i=1}^{N} \sum_{r=1}^{S} s_{ir} \log \pi_{ir}. \qquad (4)$$

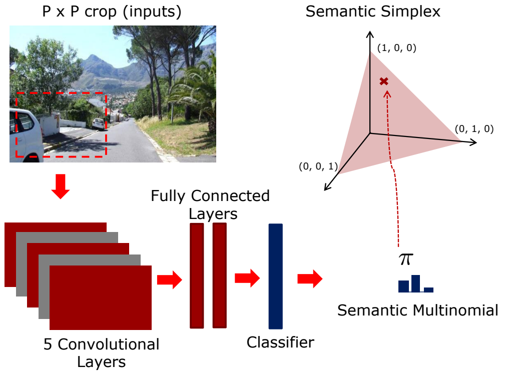

Figure 3. CNN based semantic image representation. Each image patch is mapped into an SMN π on the semantic space $\mathcal{S}$, by combination of a convolutional BoF mapping $\mathcal{F}$ and a secondary mapping $\mathcal{N}$ by the fully connected network stage. The resulting BoS is a retinotopic representation, i.e. one SMN per image patch.

Since the precise semantic labels s_i for image regions are usually not known, it is common to rely instead on the expected log-likelihood

$$E[\mathcal{L}] = \sum_{i=1}^{n} \sum_{r=1}^{S} E[s_{ir}] \log \pi_{ir} \qquad (5)$$

Using the fact that $\pi_{ir} = E[s_{ir}]$ or $P(r|x_i)$, it follows that the expected image log-likelihood is fully determined by the multinomial parameter vectors π_i. This is denoted the semantic multinomial (SMN) in [27]. They are usually computed by 1) applying a classifier, trained on the semantics of $\mathcal{V}$, to the image patches, and 2) using the resulting posterior class probabilities as SMNs π_i [21]. The process is illustrated in Figure 3 for a CNN classifier. Each patch is thus mapped into the probability simplex, which is denoted the semantic space $\mathcal{S}$ in Figure 2. The image is finally represented by the SMN collection $I = \{\pi_1, \ldots, \pi_n\}$. This is the bag-of-semantics (BoS) representation. In our implementation, we use the ImageNET classes as $\mathcal{V}$ and the ImageNET CNN [16] to estimate the SMNs π_i.

4.2. Direct FV implementation

Since a BoS is a member of the family of BoF representations, it should be possible to map it into an Euclidean space $\mathcal{D}$ through a FV embedding $\mathcal{E}$, as in Figure 1. However, because the simplex is itself not Euclidean, the operations of (1) and (3) are not directly applicable. On the other hand, it is possible to use the “Fisher recipe” with a model that is suitable for the SMN descriptors. A Dirichlet distribution is the most popular model for multinomial probability vectors [24]. Fisher gradients of a mixture of Dirichlets (DMM), are, therefore, a more natural embedding for image SMNs than the GMM-FV of (3). The log-likelihood of an image BoS $I = \{\pi_1, \ldots \pi_n\}$ under the DMM is

$$\begin{aligned} \mathcal{L} &= \log P(\{\pi_i\}_{i=1}^{n} | \{\alpha_k, w_k\}_{k=1}^{K}) & (6) \\ &= \log \prod_{i=1}^{n} \sum_{k=1}^{K} w_k \frac{\gamma\left(\sum_l \alpha_{kl}\right)}{\prod_l \gamma(\alpha_{kl})} e^{\sum_l (\alpha_{kl}-1)\log \pi_{il}}. & (7) \end{aligned}$$

[1] The results reported here are based on (3) and 128x128 image patches. They are slightly superior to those of VLAD, in our experiments

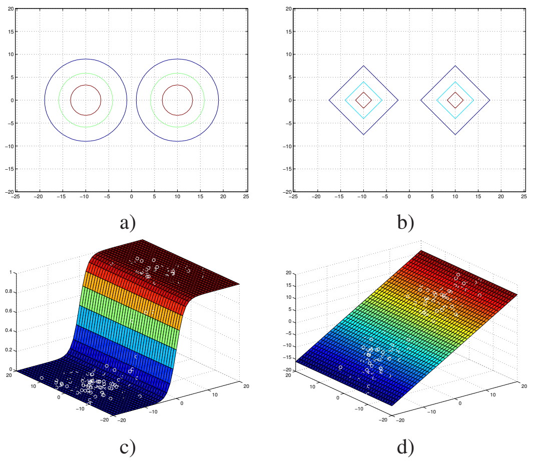

Figure 4. Top: Two classifiers in an Euclidean feature space $\mathcal{X}$, with metrics a) the L_2 or b) L_1 norms. Bottom: c) projection of a sample from a) into the semantic space $\mathcal{S}$ (only $P(y = 1|x)$ shown). The posterior surface destroys the Euclidean structure of $\mathcal{X}$ and is very similar for the Gaussian and Laplacian samples (Lapalacian omitted for brevity). d) natural parameter space mapping of c).

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

$$\begin{aligned}\mathcal{G}^I_{\alpha_k} &= \frac{1}{n}\frac{\partial \mathcal{L}}{\partial \alpha_k} \\ &= \frac{1}{n}\sum_{i=1}^{N} p(k|\pi_i)\left(\psi(\sum_l \alpha_{kl}) - \psi(\alpha_k) + \log \pi_i\right) \qquad (8)\end{aligned}$$

where $\psi(x) = \frac{\partial \gamma(x)}{\partial x}$. Using some common assumptions in the FV literature [25], we approximate the Fisher information $\mathcal{F}$ by the block diagonal matrix

$$\begin{aligned}\mathcal{F}_{lm} &= E\left[-\frac{\partial^2 \log P(\pi|\{\alpha_k, w_k\}_{k=1}^K)}{\partial \alpha_{kl}\partial \alpha_{km}}\right] \\ &\approx w_k\left(\psi'(\alpha_{kl})\delta(l,m) - \psi'(\sum_l \alpha_{kl})\right)\end{aligned} \qquad (9)$$

where $\delta(l, m) = 1$ if $l = m$. A complete derivation of $\mathcal{F}$ is given in Section 1 of the supplement. A DMM Fisher vector for image I is finally obtained from (8) and (9) as $\mathcal{F}^{-1/2}\mathcal{G}^I_{\alpha_k}$.

4.3. Limitations

While the DMM is a natural model for SMNs, our experiments show that the DMM FV does not result in an effective scene classifier (see Section 7.2). This can be attributed to a very non-Euclidean nature of the space of probability vectors. In general, the difficulty of modeling on a data space $\mathcal{X}$ depends on its topology. Most machine learning assumes vector spaces with Euclidean structure, e.g. where the natural measure of distance between examples $x_i \in \mathcal{X}$ is a metric. This is not the case for the probability simplex, which has a non-metric Kullback-Leibler divergence as its natural distance measure, and makes model learning quite difficult.

To illustrate this issue we present two binary classification problems shown in Figures 4 a) and b). In one case, the two classes are Gaussian, and in the other they are Laplacian. The class-conditional distributions of both problems are of the form $P(x|y) \propto \exp\{-d(x, \mu_y)\}$ where $Y \in \{0, 1\}$ is the class label and

$$d(x, \mu) = ||x - \mu||_p \qquad (10)$$

with $p = 1$ for Laplacian and $p = 2$ for Gaussian data. Figures 4 a) and b) show the iso-contours of the probability distributions under the two scenarios. Note that both the classifiers have very different metrics.

The posterior distribution of class $Y = 1$ is, in both cases,

$$\pi(x) = P(y = 1|x) = \sigma(d(x, \mu_0) - d(x, \mu_1)) \qquad (11)$$

where $\sigma(v) = (1 + e^{-v})^{-1}$ is the sigmoid function. Due to the non-linearity of the sigmoid mapping, the projection $x \rightarrow (\pi(x), 1 - \pi(x))$ of the samples x_i into the semantic space destroys the Euclidean structure of their original spaces $\mathcal{X}$. This is illustrated in Figure 4 c), where we show the posterior surface and the projections $\pi(x_i)$ for samples x_i of the Guassian classes of Figure 4 a). On the semantic space, the shortest path between two points is not necessarily a line. The non-linearity of the sigmoid also makes the posterior surfaces of both classification problems very similar. The posterior surface of the Laplacian problem in Figure 4 b) is visually indistinguishable from Figure 4 c) and is omitted for brevity.

The example shows two very different classifiers transforming the data into highly non-Euclidean semantic spaces that are almost indistinguishable. This suggests that modeling directly in the space of probabilities can be difficult in general. This is the most likely reason for the weakness of the DMM-FV.

5. Indirect implementation of the semantic FV

In this section, we derive an indirect implementation of the semantic Fisher vector.

5.1. Natural parameter space

For scene classification, the non-Euclidean nature of the posterior surface makes the embedding $\mathcal{E}$ of Figure 2 very difficult to learn. Note, for example, that the PCA of (1) or the Gaussian encoding of the FV in (3) make no sense for the semantic space data, since the geodesics of the posterior surface are not lines. This problem can be avoided by noting that SMNs are the parameters of the multinomial, which is a member of the exponential family of distributions

$$P_S(s; \pi) = h(s)g(\pi)\exp\left(\eta^T(\pi)T(s)\right), \qquad (12)$$

where $T(s)$ is denoted a sufficient statistic. In this family, the re-parametrization $\nu = \eta(\pi)$, makes the (log)probability distribution linear in the sufficient statistic

$$P_S(s;\nu) = h(s)g(\eta^{-1}(\nu)) \exp\left(\nu^T T(s)\right). \qquad (13)$$

This is called the *natural parameterization* of the distribution. Under this parametrization, the multinomial log-likelihood of an image BoS in (5) yields a natural parameter vector $\nu_i = \eta(E\{s_i\})$ for each patch x_i, instead of a probability vector. When the semantics are binary, the natural parameter is obtained by a logit transform $\nu = \log \frac{\pi}{1-\pi}$ of SMNs. This maps the high-nonlinear semantic space of Figure 4 c) into the linear space of Figure 4 d). Similarly, by mapping the multinomial distribution to its natural parameter space, it is possible to obtain a one-to-one transformation of the semantic space into a space with Euclidean structure. This makes the embedding $\mathcal{E}$ of Figure 2 substantially easier. In fact, it can now be implemented by the PCA in (1) and the encoding operation in (3).

5.2. Indirect FV implementation

The discussion above suggests an implementation of the semantic FV alternative to that of Section 4.2. This consists of mapping the BoS $I = \{\pi_1, \ldots \pi_n\}$ into the *natural parameter space BoS* (NP-BoS) $I = \{\nu_1, \ldots \nu_n\}$ and computing the FV of the natural parameters ν_i. As before, this is done in three steps:

1. use the PCA of (1) to map the parameters ν_i into their projection ξ_i in a lower dimensional subspace
2. learn the Gaussian mixture of (2) that best fits the low dimensional projections ξ_i
3. compute the FV of (3) for the projections ξ_i.

When compared to the direct FV implementation of Section 4.2, this implementation has the advantage of leveraging the GMM FV machinery already available in the literature. For a multinomial distribution of parameter vector $\pi = (\pi_1, \ldots, \pi_S)$ there are actually three possible natural parametrizations

$$\begin{aligned} \nu_k^{(1)} &= \log \pi_k & (14)\\ \nu_k^{(2)} &= \log \pi_k + C & (15)\\ \nu_k^{(3)} &= \log \frac{\pi_k}{\pi_S} & (16)\end{aligned}$$

where ν_k and π_k are the k^{th} entries of ν and π, respectively. The performance of these parametrizations is likely to depend on the implementation of the semantic classifier that generates the SMNs. For a discriminant classifier such as the CNN, $\nu^{(2)}$ will likely be the best parameterization. Note that, in this case, the vector of entries $\pi_k = \frac{1}{C}e^{\nu}$, is a probability vector if and only if $C = \sum_i e^{\nu_i}$. Hence, the mapping from ν to π is the softmax transformation commonly implemented at the CNN output. This implies that the CNN is learning how to discriminate the data in the natural parameter space of the multinomial distribution, which is a generalization of a natural binomial space shown in Figure 4 d). We test this assertion in Section 7.2 by comparing the parametrizations of (14)-(16) for scene classification.

6. Related work

The proposed semantic FV has relations with a number of works in the recent literature.

6.1. Square-root embedding

The semantic FV is an invariant embedding of probability vectors, based on Fisher vector encoding. The DMM-FV and the NP-BoS FV are different implementations of this idea. They provide an alternative to the popular practice of encoding square rooted probability vectors [38, 17, 6], i.e. applying the re-parametrization

$$\nu_k^{(4)} = \sqrt{\pi_k}. \qquad (17)$$

The use of the square-root is justified by differential geometric arguments in [38, 17] and as a primal embedding that induces Bhattacharya similarity between the transformed points [1]. The pooling of square-root multinomials (root-SIFT), instead of multinomials (L1-SIFT), was also shown beneficial in [6]. A comparison of this embedding with (14)-(16) and the DMM-FV is given in Section 7.2.

6.2. FVs of layer 7 activations

The proposed representation, when computed with mapping $\nu^{(2)}$ of (15), as discussed above, acts directly on the outputs of the 8^{th} layer (fc8) of ImageNET CNN [16]. In that sense, it is similar to the Fisher vectors of [10], which are computed using the activations from the fully connected 7^{th} layer (fc7). The most important difference between the two, however, is that the fc8 outputs are semantic features obtained as a result of a discriminant projection on fc7. They are, therefore, likely to be more selective. Besides their explicit semantic nature also ensures a higher level of abstraction, as a result of which they can generalize better than lower CNN layer features. We compare the two representations to validate these assertions in Section 7.3.

6.3. Fine Tuning

Beyond its success on ImageNET classification, the CNN of [16] has been shown to be highly adaptable to other classification tasks. A popular adaptation strategy, known as “fine tuning” [9], involves performing additional iterations of back-propagation on the new datasets. This is, however, an heuristic and time consuming process, which needs to be monitored carefully in order to prevent the network from over-fitting. The proposed semantic Fisher vector can also be seen as an adaptation mechanism that fully leverages the original CNN, to extract features, augmenting it with a

Fisher vector layer that enables its application to other tasks. This process is without heuristics and consumes much less time than "fine-tuning". We compare the performance of the two in Section 7.4.

6.4. The Places CNN

Recent efforts of improving scene classification have relied on a pre-trained imageNET CNN [8, 32, 10, 22]. mainly because of the superior quality of its feature responses [37]. Our work focusses on using object semantics generated by this network to obtain a high level representation for scene images. Zhou *et al*. propose a more direct approach that does not rely on the ImageNET CNN at all. They simply learn a new CNN on a large scale database of scene images known as the "Places" dataset [39]. Although the basic architecture of their Places CNN is the same as that of the ImageNET CNN, the type of features learned are very different. While the convolutional units of ImageNET CNN respond to object-like occurrences, those in Places CNN are selective of landscapes with more spatial features. The embedding of the Places CNN, therefore, produces a holistic representation of scenes that is complementary to our semantic FV. We demonstrate the effect of combining the two representations in our classification experiments.

7. Evaluation

In this section we report on a number of experiments designed to evaluate the performance of the semantic FV.

7.1. Experimental setup

All experiments were conducted on the 67 class MIT Indoor Scenes [26] and the 397 class SUN Scenes [36] datasets. The CNN features were extracted with the Caffe library [14]. For FVs, the relevant CNN features (fc7 or fc8) were extracted from local $P \times P$ image patches on a uniform grid. For simplicity, the preliminary experiments were performed with $P = 128$. A final round of experiments used multiple scale features, with $P \in \{96, 80, 128, 160\}$. For all GMM-FVs the local features were first reduced to 500 dimensions, using a PCA, and then pooled using (3) and a 100 component mixture. The DMM-FV of Section 4.2 was learned with a 50 component mixture on the 1,000 dimensional SMN space. As is common in the literature,all Fisher vectors were power normalized and L_2 normalized. This resulted in DMM and GMM FVs of size of 50000, dimensions of which were further reduced to 4096, by PCA. In some experiments, we also evaluate classifiers based on fc7 and SMN features extracted globally, as in [8]. The global fc7 features were square-rooted and L_2 normalized, whereas the global SMNs were simply square rooted. Scene classifiers trained on all image representations were implemented with a linear SVM.

Table 3. Comparison of different Fisher vector encoding techniques over SMNs.

Method	MIT Indoor	SUN
DMM-FV	58.8	40.86
$\nu^{(1)}$-FV	67.7	49.86
$\nu^{(2)}$-FV	**68.5**	**49.86**
$\nu^{(3)}$-FV	67.6	48.81
$\nu^{(4)}$-FV	58.95	40.6
π-FV	55.3	36.87

7.2. Direct vs. indirect Semantic FVs

We start with a comparison of the direct and indirect implementations of the semantic Fisher vectors. The former is based on the DMM-FV of Section 4.2, the latter on the parameter mappings of (14)-(17), which are denoted $\nu^{(i)}$-FV. An additional benchmark is introduced, denoted as π-FV, which is a GMM based Fisher vector (1)-(3) applied directly on the SMNs. We conduct this experiment on a single training/test split of MIT Indoors and SUN datasets and report the accuracies in Table 3. All the indirect implementations of the semantic FV perform substantially better than the remaining methods, with up to a 10 points gain. The poor performance of the DMM-FV reflects the previously noted difficulties of modeling on the simplex. The square-root projection onto a great circle of (17) does not fare better. The linear mapping π-FV has the overall worst performance. Among the indirect implementations of the semantic FV, (15) achieves the best results, although the differences are subtle. Given these results, we adopt the indirect implementation of the semantic FV, with the reparametrization $\nu^{(2)}$ of (15) in the remaining experiments. This is simply denoted as "the semantic FV."

7.3. The role of invariance

To test the hypothesis that the semantic FV is both more discriminant and invariant than FVs extracted from lower network layers, we compared its performance with that of the fc7 FV of [10]. In this experiment, the CNN features were extracted at multiple scales (globally as well as from patches of size 80, 96, 128 and 160 pixels). Table 2 shows the results of the sematic FV (denoted fc8) and the fc7 FV. Several remarks are worth making, in light of previous reports on similar experiments [9, 8, 10]. First, when compared to the approach of extracting CNN features globally [8], the localized representations have far better performance. Second, while fc7 features extracted globally are known to perform poorly [10], the use of global 8^{th} layer features leads to even worse performance. This could suggest the conclusion that layer 8 somehow extracts "worse features" for scene classification. The remaining columns, however, clearly indicate otherwise. When extracted locally, semantic descriptors are highly effective, achieving a gain of up to 3 points with respect to the fc7 features. The gap in performance between the localized and global semantic descriptors is explained by the localized nature of

Table 2. Impact of semantic feature extraction at different scales.

Dataset	Feat	full img	160x160	128x128	96x96	80x80	Best 3	Best 4
Indoor	fc8	48.5	**66.6**	**68.5**	**67.8**	**67.38**	**71.24**	**72.86**
	fc7	**59.5**	64.7	65.1	65.4	65.37	68.8	69.7
SUN	fc8	32.6	47.5 ± 0.71	**49.61 ± 0.22**	**50.03 ± 0.24**	**49.39 ± 0.28**	**53.24 ± 0.16**	**54.4 ± 0.3**
	fc7	**43.76**	48.08 ± 0.52	48.3 ± 0.63	48.46 ± 0.25	47.44 ± 0.1	51.8 ± 0.5	53.0 ± 0.4

scene semantics, which vary from patch to patch. A global semantic descriptors is just not expressive enough to capture this diversity. Third, recent arguments for the use of intermediate CNN features should be revised. On the contrary, the results of the table support the conclusion that these features are both less discriminant and invariant than semantic descriptors. When combined with a proper encoding, such as the semantic FV, the latter achieve the best scene classification results.

Finally, to ensure that the gains of the semantic FV are not just due to the use of the transformation of (15), we applied the transformation to the fc7 features as well. Rather than a gain, this resulted in a substantial performance decrease (58% compared to the 65.1% of the fc7-FV on MIT Indoors at patch size 128). This was expected, since the natural parameter space arguments do not apply in this case.

7.4. Comparison to the state of the art

Concatenating Fisher vectors of fc7 features computed at multiple patch scales was shown to produce substantial gains in [10]. We implemented this strategy for both the fc7-FV and the semantic FV, with the results shown in Table 2. Combining the fc7-FVs at three patch scales resulted in classification accuracies of 68.8% on MIT Indoor and 51.8% on SUN. While this is a non-trivial improvement over any of the single-scale classifiers, the concatenation of semantic-FVs at 3 scales produced even better results (accuracies of 71.24% on MIT Indoor and 53.24% on SUN). Similar gains were observed when using 4 patch scales, as reported in the table.

A comparison of our multiscale semantic FV with other leading representations derived from the ImageNET CNNs is shown in table 4. As expected, the pioneering DeCaf [8] representation is vastly inferior to all other methods since it describes complex scene images with a globally extracted descriptor using an object CNN [16]. Among techniques that rely on local feature extraction are the proposals of Liu *et al.* [22] and Razavian *et al.* [32]. The scene representation in [22] is a sparse code derived from the 6^{th} layer activations (fc6) of the CNN. Razavian *et al.* [32], use features from the penultimate layer of the OverFeat network [31] extracted from coarser spatial scales.Since, the features used in both [22] and [32] lack the invariance of semantics, their classifiers are easily outperformed by our semantic FV classifier. We also compare with the technique referred to as fine-tuning [9] which adapts the imageNET CNNs directly to the task of scene classification. The process requires a few tens of thousands of back-propagation iterations on the scene dataset of interest and lasts about 5-10 hours on a single GPU. The resulting classifier, however, is significantly worse than our semantic FV classifier.

Table 4. Comparison with the state-of-the-art methods using ImageNET trained features. *-Indicates our implementation.

Method	MIT Indoor	MIT SUN
fc8-FV (Our)	**72.86**	**54.4 ± 0.3**
fc7-FV [10]*	69.7	53.0 ± 0.4
fc7-VLAD [10]	68.88	51.98
ImgNET finetune	63.9	48.04 ± 0.19
OverFeat + SVM [32]	69	-
fc6 + SC [22]	68.2	-
DeCaF [8]*	59.5	43.76

Table 5. Comparison with a CNN trained on Scenes [39]

Method	MIT Indoor	MIT SUN
ImgNET fc8-FV (Our)	72.86	54.4 ± 0.3
Places fc7 [39]	68.24	54.34 ± 0.14
Combined	**79.0**	**61.72 ± 0.13**

An alternative to using pre-trained object classification CNNs [16, 31] for scenes is to learn a CNN directly on a large scale scene dataset. This was recently performed by Zhou *et al.* using a 2 million image Places dataset [39]. Table 5 indicates a comparison between a scene representation obtained with the Places CNN and our ImageNET based semantic FV. The results of semantic FV are slightly better that theirs on the Indoor scenes dataset, whereas, on SUN, both the descriptors perform comparably. More importantly, a simple concatenation of the two produces a gain of almost 7% in accuracy on both datasets, indicating that the embeddings are, in fact, complimentary. These results are, to the best of our knowledge, state-of-the-art on scene classification.

8. Conclusions

In this paper we discussed the benefits of modeling scene images as bags of object semantics from an ImageNET CNN instead of its lower layer activations. To leverage the superior quality of semantic descriptors, we propose an effective approach to summarize them with a Fisher vector, which is non-trivial. The semantic FV provides a better classification architecture than an FV of low-level features or a even fine-tuned classifier. When combined with features from a scene classification CNN, our semantic FV produces state-of-the-art results.

9. Acknowledgements

This work was partially funded by NSF awards IIS-1208522 and CCF-0830535 and a gift from Yahoo! Faculty Research Excellence Program (FREP).

References

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[2] A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[3] A. Bergamo and L. Torresani. Classemes and other classifier-based features for efficient object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2014. 1, 3

[4] A. Bergamo, L. Torresani, and A. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2088–2096. 2011. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society. 1

[6] J. Delhumeau, P. H. Gosselin, H. Jégou, and P. Pérez. Revisiting the vlad image representation. In *ACM Multimedia*, pages 653–656, 2013. 6

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255, June 2009. 3

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition, 2014. 1, 2, 3, 7, 8

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2, 6, 7, 8

[10] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014*, volume 8695 of *Lecture Notes in Computer Science*, pages 392–407. Springer International Publishing, 2014. 1, 2, 4, 6, 7, 8

[11] G. Griffin, A. Holub, and P. Perona. The caltech-256. Technical report, caltech technical report, 2006. 3

[12] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 487–493, Cambridge, MA, USA, 1999. MIT Press. 1

[13] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3304–3311, jun 2010. 4

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 3, 7

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1, 2, 3, 4, 6, 8

[17] R. Kwitt, N. Vasconcelos, and N. 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. 1, 2, 3, 6

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 1

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. 1

[20] L.-J. Li, H. Su, Y. Lim, and F.-F. Li. Object bank: An object-level image representation for high-level visual recognition. *International Journal of Computer Vision*, 107(1):20–39, 2014. 1, 3

[21] W. Li and N. Vasconcelos. Recognizing activities by attribute dynamics. In *Advances in Neural Information Processing Systems*, pages 1115–1123, 2012. 4

[22] L. Liu, C. Shen, L. Wang, A. Hengel, and C. Wang. Encoding high dimensional local features by sparse coding based fisher vectors. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1143–1151. Curran Associates, Inc., 2014. 2, 7, 8

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints, 2003. 1

[24] T. P. Minka. Estimating a dirichlet distribution. Technical report, 2000. 4

[25] F. Perronnin, J. Sánchez, and T. 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. 1, 2, 3, 5

[26] A. Quattoni and A. Torralba. Recognizing indoor scenes. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 0:413–420, 2009. 3, 7

[27] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *Multimedia, IEEE Transactions on*, 9(5):923–938, 2007. 1, 4

[28] N. Rasiwasia and N. Vasconcelos. Scene classification with low-dimensional semantic spaces and weak supervision. In *IEEE CVPR*, 2008. 3

[29] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *Pattern Analysis and Machine Intel-*

ligence, IEEE Transactions on, 34(5):902–917, May 2012. 1

[30] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013. 1, 3

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. 2, 8

[32] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014. 2, 3, 7, 8

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

[34] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision (ECCV)*, pages 776–789, Sept. 2010. 1, 3

[35] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *Int. J. Comput. Vision*, 72(2):133–157, Apr. 2007. 1

[36] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3485–3492, 2010. 7

[37] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833, 2014. 2, 7

[38] D. Zhang, X. Chen, and W. S. Lee. Text classification with kernels on the multinomial manifold. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 266–273, New York, NY, USA, 2005. ACM. 6

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014. 1, 2, 7, 8