Unsupervised Learning of Semantics of Object Detections for Scene Categorization

Grégoire Mesnil, Salah Rifai, Antoine Bordes, Xavier Glorot, Yoshua Bengio and Pascal Vincent

Abstract Classifying scenes (e.g. into "street", "home" or "leisure") is an important but complicated task nowadays, because images come with variability, ambiguity, and a wide range of illumination or scale conditions. Standard approaches build an intermediate representation of the global image and learn classifiers on it. Recently, it has been proposed to depict an image as an aggregation of its contained objects: the representation on which classifiers are trained is composed of many heterogeneous feature vectors derived from various object detectors. In this paper, we propose to study different approaches to efficiently learn contextual semantics out of these object detections. We use the features provided by Object-Bank [24] (177 different object detectors producing 252 attributes each), and show on several benchmarks for scene categorization that careful combinations, taking into account the structure of the data, allows to greatly improve over original results (from +5 to +11 %) while drastically reducing the dimensionality of the representation by 97 % (from 44,604 to 1,000). We also show that the uncertainty relative to object detectors hampers the use of external semantic knowledge to improve detectors combination, unlike our unsupervised learning approach.

Keywords Unsupervised learning · Transfer learning · Deep learning · Scene categorization · Object detection

1 Introduction

Automatic scene categorization is crucial for many applications such as content-based image indexing [37] or image understanding. This is defined as the task of assigning images to predefined categories ("office", "sailing", "mountain", etc.).

G. Mesnil (✉) · S. Rifai · X. Glorot · Y. Bengio · P. Vincent
LISA, Université de Montréal, Québec, Canada
e-mail: gregoire.mesnil@gmail.com

G. Mesnil
LITIS, Université de Rouen, Rouen, France

A. Bordes
CNRS - Heudiasyc UMR 7253, Université de Technologie de Compiègne, Compiègne, France

A. Fred and M. De Marsico (eds.), *Pattern Recognition Applications and Methods*,
Advances in Intelligent Systems and Computing 318,
DOI 10.1007/978-3-319-12610-4_13

Classifying scene is complicated because of the large variability of quality, subject and conditions of natural images which lead to many ambiguities w.r.t. the corresponding scene label.

Standard methods build an intermediate representation before classifying scenes by considering the image as a whole [10, 28, 38, 40]. In particular, many such approaches rely on power spectral information, such as magnitude of spatial frequencies [28] or local texture descriptors [10]. They have shown to perform well in cases where there are large numbers of scene categories.

Another line of work conveys promising potential in scene categorization. First applied to object recognition [9], attribute-based methods have now proved to be effective for dealing with complex scenes. These models define high-level representations by combining semantic lower-level elements, e.g., detection of object parts. A precursor of this tendency for scenes was an adaptation of pLSA [15] to deal with "visual words" proposed by [5]. An extension of this idea consists in modeling an image based on its content i.e., its objects [7, 24]. Hence, the Object-Bank (OB) project [25] aims at building high-dimensional over-complete representations of scenes (of dimension 44,604) by combining the outputs of many object detectors (177) taken at various poses, scales and positions in the original image (leading to 252 attributes per detector). Experimental results indicate that this approach is effective since simple classifiers such as Support Vector Machines trained on their representations achieve state-of-the-art performance. However, this approach suffers from two flaws: (1) curse of dimensionality (very large number of features) and (2) individual object detectors have a poor precision (30 % at most). To solve (1), the original paper proposes to use structured norms and group sparsity to make best use of the large input. Our work studies new ways to combine the very rich information provided by these multiple detectors, dealing with the uncertainty of the detections. A method designed to select and combine the most informative attributes would be able to carefully manage redundancy, noise and structure in the data, leading to better scene categorization performance.

Hence, in the following, we propose a sequential 2-steps strategy for combining the feature representations provided by the OB object detectors on which the linear SVM classifier is destined to be trained for categorizing scenes. The first step adapts Principal Components Analysis (PCA) to this particular setting: we show that it is crucial to take into account the structure of the data in order for PCA to perform well. The second one is based on *Deep Learning*. Deep Learning has emerged recently (see [3] for a review) and is based on neural network algorithms able to discover data representations in an unsupervised fashion [2, 14, 18, 19, 32]. We propose to use this ability to combine multiple detector features. Hence, we present a model trained using Contractive Auto-Encoders [33, 34], which have already proved to be efficient on many image tasks and has contributed to winning a transfer learning challenge [26].

We validate the quality of our models in an extensive set of experiments in which several setups of the sequential feature extraction process are evaluated on benchmarks for scene classification [21, 23, 31, 41]. We show that our best results substantially outperform the original methods developed on top of OB features, while

producing representations of much lower dimension. The performance gap is usually large, indicating that advanced combinations are highly beneficial. We show that our method based on dimensionality reduction followed by deep learning offers a flexibility which makes it able to benefit from semi-supervised and transfer learning.

2 Scene Categorization with Object-Bank

Let us begin by introducing the approach of the OB project [24]. First, the 177 most useful (or frequent) objects were selected from popular image datasets such as LabelMe [35], ImageNet [6] and Flickr. For each of these 177 objects, a specific detector, existing in the literature [11, 16], was trained. Every detector is composed of 2 *root filters* depending on the pose, each one coming with its own deformable pattern of parts, e.g., there is one root filter for the front-view of a bike and one for the side-view. These $354 = 177 \times 2$ part-based filters (each composed by a root and its parts) are used to produce features of natural images. For a given image, a filter is convolved at 6 different scales. At each scale, the max-response among $21 = 1 + 4 + 16$ positions (whole image, quadrants, quadrants within each quadrant) is kept, producing a response map of dimension $126 = 6 \times 21$. All 2×177 maps are finally concatenated to produce an over-complete representation $x \in \mathbb{R}^{44,604}$ of the original image.

In the original OB paper [24], classifiers for scene categorization are learned directly on these feature vectors of dimension 44,604. More precisely, C classifiers (Linear SVM or Logistic Regression) are trained in a 1-versus-all setting in order to predict the correct scene category $y_{\text{category}}(x)$ among C different categories. Various strategies using structured sparsity with combinations of ℓ_1/ℓ_2 norms have been proposed to handle the very large input.

3 Unsupervised Feature Learning

The approach of OB for the task of scene categorization, based on specific object detectors, is appealing since it works well in practice. This suggests that a scene is better recognized by first identifying basic objects and then exploiting the underlying semantics in the dependencies between the corresponding detectors.

However, it appears that none of the individual object detectors reaches a recognition precision of more than 30 %. Hence, one may question whether the ideal view that inspired this approach (and expressed above) is indeed the reason of OB's success. Alternatively, one may hypothesize that the 44,604 OB features are more useful for scene categorization because they represent high level statistical properties of images than because they precisely report the absence/presence of objects—see Fig. 1. OB tried structured sparsity to handle this feature selection but there may be other ways—simpler or not.

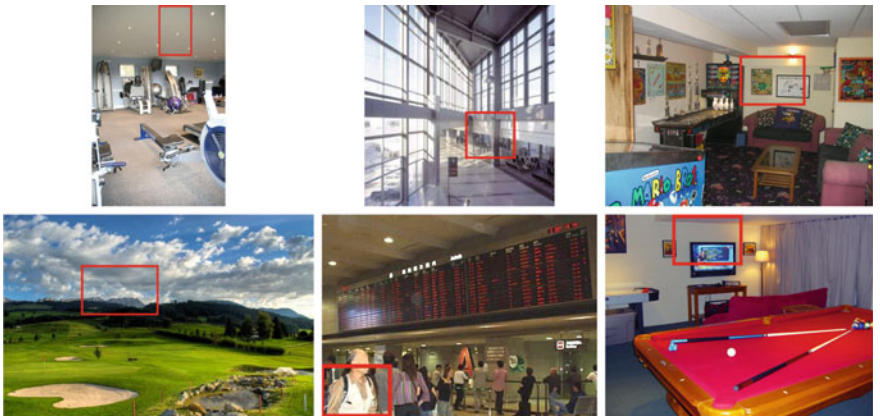

Fig. 1 *Left* Cloud *Middle* Man *Right* Television. *Top* False Detections *Bottom* True Detections. Images from **SUN** [41] for which we compute the OB representation and display the bounding box around the average position of various objects detectors. For instance, the *television* detector can be viewed either as a *television* detector or a *rectangle* shape detector i.e. high-order statistical properties of the image

This paper investigates several ways of learning higher-level features **on top of** the high dimensional representation provided by OB, expecting that capturing further structure may improve categorization performance. Our approach employs *unsupervised feature learning/extraction* algorithms, i.e. generic feature extraction methods which were not developed specifically for images. We will consider both standard Principal Component Analysis and Contractive Auto-Encoders [33, 34]. The latter is a recent machine learning method which has proved to be a robust feature extraction tool.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) [17, 30] is the most prevalent technique for linear dimensionality reduction. A PCA with k components finds the k orthonormal directions of projection in input space that retain most of the *variance* of the training data. These correspond to the eigenvectors associated with the leading eigenvalues of the training data's covariance matrix. Principal components are ordered, so that the first corresponds to the direction along which the data varies the most (largest eigenvalue), etc…

Since we will consider an auto-encoder variant (presented next), we should mention here a well-known result: a linear auto-encoder with k hidden units, trained to minimize squared reconstruction error, will learn projection directions that span the same *subspace* as a k component PCA [1]. However the regularized non-linear

auto-encoder variant that we consider below is capable of extracting qualitatively different, and usually more useful, nonlinear features.

3.2 Contractive Auto-Encoders

Contractive Auto-Encoders (CAEs) [33, 34] are among the latest developments in a line of machine learning research on nonlinear feature learning methods, that started with the success of Restricted Boltzmann Machines [14] for pre-training deep networks, and was followed by other variants of auto-encoders such as sparse [13, 19, 32] and denoising auto-encoders [39]. It was selected here mainly due to its practical ease of use and recent empirical successes.

Unlike PCA that decomposes the input space into leading *global* directions of variations, the CAE learns features that capture local directions of variation (in some regions of input space). This is achieved by penalizing the norm of the Jacobian of a latent representation $h(x)$ with respect to its input x at training samples. In [34], authors show that the resulting features provide a local coordinate system for a low dimensional manifold of the input space. This corresponds to an atlas of charts, each corresponding to a different region in input space, associated with a different set of active latent features. One can think about this as being similar to a mixture of PCAs, each computed on a different set of training samples that were grouped together using a similarity criterion (and corresponding to a different input region), but without using an independent parametrization for each component of the mixture, i.e., allowing to generalize across the charts, and away from the training examples.

In the following, we summarize the formulation of the CAE as a regularized extension of a basic Auto-Encoder (AE). In our experiments, the parametrization of this AE consists in a non-linear encoder or latent representation h of m hidden units with a linear decoder or reconstruction g towards an input space of dimension d.

Formally, the latent variables are parametrized by:

$$h(x) = s(Wx + b_h), \tag{1}$$

where s is the element-wise logistic sigmoid $s(z) = \frac{1}{1+e^{-z}}$, $W \in \mathcal{M}_{m\times d}(\mathbb{R})$ and $b_h \in \mathbb{R}^m$ are the parameters to be learned during training. Conversely, the units of the decoder are linear projections of $h(x)$ back into the input space:

$$g(h(x)) = W^T h(x). \tag{2}$$

Using mean squared error as the reconstruction objective and the L2-norm of the Jacobian of h with respect to x as regularization, training is carried out by minimizing the following criterion by stochastic gradient descent:

$$\mathcal{J}_{\mathrm{CAE}}(\Theta) = \sum_{x\in\mathcal{D}} \|x - g(h(x))\|^2 + \lambda \sum_{i=1}^{m}\sum_{j=1}^{d} \left|\frac{\partial h_i}{\partial x_j}(x)\right|^2, \tag{3}$$

where $\Theta = \{W, b_h\}$, $\mathcal{D} = \{x^{(i)}\}_{i=1,\ldots,n}$ corresponds to a set of n training samples $x \in \mathbb{R}^d$ and λ is a hyper-parameter controlling the level of contraction of h. A notable difference between CAEs and PCA is that features extracted by CAEs are non-linear w.r.t. the inputs, so that multiple layers of CAEs can be usefully composed (stacked), whereas stacking linear PCAs is pointless.

4 Extracting Better Features with Advanced Combination Strategies

In this work, we study two different sub-structures of OB. We consider the *pose* response defined by the output of only one part-based filter at all positions and scales, and the *object* response which is the concatenation of all *pose* responses associated to an object. Combination strategies are depicted in Fig. 2.

4.1 Simplistic Strategies: Mean and Max Pooling

The idea of pooling responses at different locations or poses has been successfully used in Convolutional Neural Networks such as LeNet-5 [22] and other visual processing [36] architectures inspired by the visual cortex.

Here, we pool the 252 responses of each object detector into one component (using the mean or max operator) leading to a representation of size $177 = 44{,}604/252$. It corresponds to the mean/max over the object responses at different scales and locations. One may view the object max responses as features encoding absence/presence of objects while discarding all the information about the detector's positions.

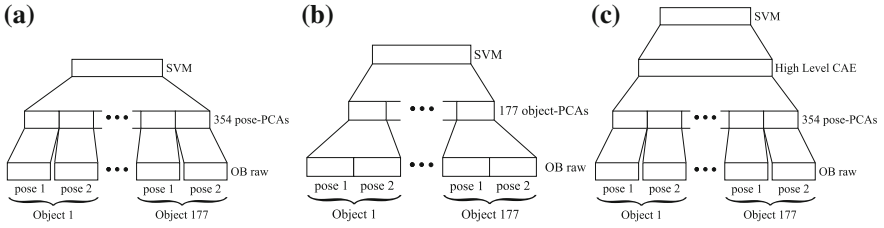

Fig. 2 Different Combination Strategies (**a**) and (**b**) *pose* and *object* PCAs (**c**) high-level CAE: *pose*-PCA as dimensionality reduction technique in the first layer and a CAE stacked on top. We denote it high-level because it can learn *context information* i.e. plausible joint appearance of different objects

4.2 Combination Strategies with PCA

PCA is a standard method for extracting features from high dimensional input, so it is a good starting point. However, as we find in our experiments, exploiting the particular structure of the data, e.g., according to poses, scales, and locations, can yield to improved results.

Whole PCA. An ordinary PCA is trained on the raw output of OB ($x \in \mathbb{R}^{44,604}$) without looking for any structure. Given the high-dimensionality of OB's representation, we used the Randomized PCA algorithm of the scikits toolbox.[1]

Pose-PCA. Each of the two *poses* associated with each *object* detector is considered independently. This results in $354 = 2 \times 177$ different PCAs, which are trained on *pose* outputs ($x \in \mathbb{R}^{126}$)—see Fig. 2.

Object-PCA. Only each *object* response ($x \in \mathbb{R}^{252}$) is considered separately, therefore 177 PCAs are trained in total. It allows the model to capture variations among all *pose* responses at various scales and positions—see Fig. 2.

Note that, in all cases, whitening the PCA (i.e. dividing each eigenvector's response by the corresponding squared root eigenvalue) performs very poorly. For post-processing, the PCA outputs $\tilde{x}$ are always normalized: $\tilde{x} \leftarrow (\tilde{x} - \mu)/\sigma$ according to mean μ and the deviation σ of the whole, per *object* or per *pose* PCA outputs. Thereby, we ensure contributions from all *objects* or *poses* to be in the same range. The number of components in all cases has been selected according to the classification accuracy estimated by 5-fold cross-validation.

4.3 Improving upon PCA with CAE

Due to hardware limitations and high-dimensional input, we could not train a CAE on the whole OB output ("whole CAE"). However, we address this problem with the sequential feature extraction steps below.

To overcome the tractability problem that forbids a CAE to be trained on the whole OB output, we preprocess it by using the *pose*-PCAs as a dimensionality reduction method. We keep only the 5 first components of each *pose*. Given this low-dimensional representation (of dimension $1,770$), we are able to train a CAE—see Fig. 2. The CAE has a global view of all object detectors and can thus learn to capture *context information*, defined by the joint appearance of combinations of various objects. Moreover, instead of using an SVM on top of the learned representations, we can use a Multi-Layer Perceptron whose weights would be initialized by those of this CAE. This setting is where the CAE has shown to perform best in practice [33].

[1] Available from http://scikits.appspot.com/.

5 Experiments

5.1 Datasets

We evaluate our approach on 3 scene datasets, cluttered indoor images (MIT Indoor Scene), natural scenes (15-Scenes), and event/activity images (UIUC-Sports). Images from a large scale scene recognition dataset (SUN-397 database) have also been used for unsupervised learning.

- **MIT Indoor** is composed of 67 categories and, following [24, 31], we used 80 images from each category for training and 20 for testing.
- **15-Scenes** is a dataset of 15 natural scene classes. According to [21], we used 100 images per class for training and the rest for testing.
- **UIUC-Sports** contains 8 event classes. We randomly chose 70 / 60 images for our training / test set respectively, following the setting of [23, 24].
- **SUN-397** contains a full variety of 397 well sampled scene categories (100 samples per class) composed of 108,754 images in total.

5.2 Tasks

We consider 3 different tasks to evaluate and compare the considered combination strategies. In particular, various supervision settings for learning the CAE are explored. Indeed, a great advantage of this kind of method is that it can make use of vast quantities of unlabeled examples to improve its representations. We thus illustrate this by proposing experiments in which the CAE has been trained in supervised or in semi-supervised way and also in a transfer context.

MIT Indoor (*plain*). Only the official training set of the MIT Indoor scene dataset (5,360 images) is used for unsupervised feature learning. Each representation is evaluated by training a linear SVM on top of the learned features.

MIT + SUN (*semi-supervised*). This task, like the previous one, uses the official train/test split of the MIT Indoor scene dataset for its supervised training and evaluation of scene categorization performance. For the initial unsupervised feature extraction however, we augmented the MIT Indoor training set with the whole dataset of images from SUN-397 (108,754 images). This yields a total of 123,034 images for unsupervised feature learning and corresponds to a *semi-supervised* setting. Our motivation for adding scene images from SUN, besides increasing the number of training samples, is that on MIT Indoor, which contains only indoor scenes, OB detectors specialized on outdoor objects would likely be mostly inactive (as a sailboat detector applied on indoor scenes) and irrelevant, introducing an harmful noise in the unsupervised feature learning. As SUN is composed of a wide range of indoor and outdoor scene images, its addition to MIT Indoor ensures that each detector

meaningfully covers its whole range of activity (having a "balanced" number of positives/negatives detections through the training set) and the feature extraction methods can be efficiently trained to capture it.

One may object that training on additional images does not provide a fair comparison w.r.t. the original OB method. Nevertheless, we recall that (1) the supervised classifiers do not benefit from these additional examples and (2) object detectors which are the core of OB representations (and all detector-based approaches) have also obviously been trained on *additional* images.

UIUC-Sports and 15-Scenes (*transfer*). We would also like to evaluate the discriminative power of the various representations learned on the MIT + SUN dataset, but on new scene images and categories that were *not* part of the MIT + SUN dataset. This might be useful in case other researchers would like to use our compact representation on a different set of images. Using the representation output by the feature extractors learned with MIT+SUN, we train and evaluate classifiers for scene categorization on images from UIUC-Sports and 15-Scenes (not used during unsupervised training). This corresponds to a *transfer learning* setting for the feature extractors.

5.3 SVMs on Features Learned with Each Strategy

In order to evaluate the quality of the features generated by each strategy, a linear SVM is trained on the features extracted by each combination method. We used LibLinear [8] as SVM solver and chose the best C according to 5-fold cross-validation scheme. We compare accuracies obtained by features provided by all considered combination methods against the original OB performances [24]. Results obtained with SVM classifiers on all MIT-related tasks are displayed in Table 1 and those concerning UIUC and 15-scenes in Table 2.

The simplistic strategy *object* mean-pooling performs surprisingly well on all datasets and tasks whereas *object* max-pooling obtained the worst results. It suggests that taking the mean response of an object detector across various scales and positions is actually meaningful compared to consider presence/absence of objects as max-pooling does.

On MIT and MIT + SUN, *object* or *pose* PCAs reach almost the same range of performance slightly above the current state-of-the-art performances [29], except for whole-PCA which performs poorly: one must consider the structure of OB to combine features efficiently. In the experiments, keeping the 10 (resp. 15) first principal components gave us the best results for pose-PCA (resp. object-PCA).

Besides, Table 3 shows that both PCAs and PCA + CAE allow a huge reduction of the dimension of the OB feature representation.

Results obtained for the UIUC-Sports and 15-Scenes transfer learning tasks are displayed in Table 2. Representations learned on MIT + SUN generalize quite well and can be easily used for other datasets even if images from those datasets have not been seen at all during unsupervised learning.

Table 1 MIT Indoor

	MIT (*plain*) (%)	MIT+SUN (*semi-supervised*) (%)
object-MAX + SVM	24.3	–
object-MEAN + SVM	41.0	–
whole-PCA + SVM	40.2	–
object-PCA + SVM	42.6	46.1
pose-PCA + SVM	40.1	46.0
pose-PCA + MLP	42.9	46.3
pose-PCA + CAE (MLP)	**44.0**	**49.1**
Object Bank + SVM	37.6	–
Object Bank + rbf-SVM	37.7	–
DPM + Gist + SP	43.1	–
Improvement w.r.t. Object Bank	+6.4	+11.5

Results are reported on the official split [31] for all combination strategies described in Sect. 4. Only the unsupervised feature learning strategies (PCA and CAE based) *can* benefit from the addition of unlabeled scenes from SUN. Object Bank + SVM refers to the original system [24] and DPM + Gist + SP [29] corresponds to the state-of-the-art method on MIT Indoor

Table 2 UIUC Sports and 15-Scenes

	UIUC-Sports (%)	15-SCENES (%)
object-MAX + SVM	67.23 ± 1.29	71.08 ± 0.57
object-MEAN + SVM	81.88 ± 1.16	83.17 ± 0.53
object-PCA + SVM	83.90 ± 1.67	85.58 ± 0.48
pose-PCA + SVM	83.81 ± 2.22	85.69 ± 0.39
pose-PCA + MLP	84.29 ± 2.23	84.93 ± 0.39
pose-PCA + CAE (MLP)	$\mathbf{85.13 \pm 1.07}$	$\mathbf{86.44 \pm 0.21}$
Object Bank + SVM	78.90	80.98
Object Bank + rbf-SVM	78.56 ± 1.50	83.71 ± 0.64
Improvement w.r.t. OB	+6.23	+5.46

Results are reported for 10 random splits and compared to the original OB results [24]—Object Bank + SVM—on one single split

Table 3 Dimensionality reduction

Object-Bank	Pooling	*whole*-PCA	*object*-PCA	*pose*-PCA	*pose*-PCA+CAE
44,604	177	1,300	2,655	1,770	1,000

Dimension of representations obtained on MIT Indoor. The *pose*-PCA + CAE produces a compact and powerful combination

5.4 Deep Learning with Fine Tuning

Previous work [20] on Deep Learning generally showed that the features learned through unsupervised learning could be improved upon by fine-tuning them through a supervised training stage. In this stage (which follows the unsupervised pre-training stage), the features and the classifier on top of them are together considered to be a supervised neural network, a Multi-Layer Perception (MLP) whose hidden layer is the output of the trained features. Hence we apply this strategy to the *pose* PCA + CAE architecture, keeping the PCA transformation fixed but fine-tuning the CAE and the MLP altogether. These results are given at the bottom of Tables 1 and 2. The MLP are trained with early stopping on a validation set (taken from the original training set) for 50 epochs.

This yields 44.0 % test accuracy on plain MIT and 49.1 % on MIT + SUN: this allows to obtain state-of-the-art performance, with or without semi-supervised training of the CAEs, even if these additional examples are highly beneficial. As a check, we also evaluate the effect of the unsupervised pre-training stage by completely skipping it and only training a regular supervised MLP of 1,000 hidden units on top of the PCA output, yielding a worse test accuracy of 42.9 % on MIT and 46.3 % on MIT+SUN. This improvement with fine-tuning on labeled data is a great advantage for CAE compared to PCA. Fine-tuning is also beneficial on UIUC-Sports and 15-Scenes. On both datasets, this leads to an improvement of +6 and +5 % w.r.t the original system.

Finally, we trained a non-linear SVM (with rbf kernel) to verify whether this gap in performances was simply due to the replacement of a linear classifier (SVM) by a non-linear one (MLP) or to the detectors' outputs combination. The poor results of the rbf-SVM (see Tables 1 and 2) suggests that the careful combination strategies are essential to reach good performance (Table 4).

Table 4 *Context semantics*

Context Semantics learned by the CAE
Sailboat, rock, tree, coral, blind
Roller coaster, building, rail, keyboard, bridge
Sailboat, autobus, bus stop, truck, ship
Curtain, bookshelf, door, closet, rack
Soil, seashore, rock, mountain, duck
Attire, horse, bride, groom, bouquet
Bookshelf, curtain, faucet, screen, cabinet
Desktop computer, printer, wireless, computer screen

Names of the detectors corresponding to the highest weights of 8 hidden units of the CAE. These hidden units will fire when those objects will be detected altogether

5.5 Use of External Semantic Information for Re-ranking

WordNet's [27] semantic structure provides an easy way to measure word similarities. We assume that closely related objects detectors (according to WordNet) should fire together and could be grouped in order to build semantically meaningful features. E.g. by grouping the output of *ship*, *sea* and *sun* into a single feature, the combination's output might be useful for classifying the "sailing" scene category.

In our experiments, we used the lesk distance in WordNet to extract the neighbors of each detector's name. Some examples are depicted in Table 5. Afterwards, given the score $s(x) \in \mathbb{R}^{177}$ obtained with the mean-pooling strategy from the original OB representation $x \in \mathbb{R}^{44,604}$, we performed the following Re-Ranking operation:

$$s_i'(x) = \sum_{j=1}^{177} s_j(x)\gamma^{R(i,j)} \quad \text{for} \quad i = 1, \ldots, 177 \tag{4}$$

where $\gamma \in [0, 1]$ is a decay hyper-parameter tuned on a validation set. $R(i, j)$ corresponds to the rank of the object j among the neighbors of object i according to the lesk metric ($R(i, i) = 0$). Results are presented in Table 6. The relatively small improvement brought by WordNet illustrates the fact that the poor intrinsic quality of the object detectors prevents any use of external semantic resource to improve their combination.

Table 5 *WordNet semantics* Names of the detectors and their top-ranked neighbors according to the lesk distance computed from WordNet

Rank	Bus	Lion	Laptop
1.	Car	Tree	Baggage
2.	Ship	Dog	Desktop computer
3.	Truck	Bird	Computer
4.	Aircraft	Horse	Bed
5.	Train	Computer	Door

Table 6 *Re-Ranking* Results are reported on the official split [31]

object-MEAN+SVM	MIT (*plain*)
w/o Re-Ranking	41.03 %
with Re-Ranking	41.52 %

Object-mean+SVM refers to the mean-pooling strategy with and w/o the Re-Ranking transformation

6 Discussion

In this work, we add one or more levels of trained representations on top of the layer of object and part detectors (OB features) that have constituted the basis of very promising trend of approach for scene classification [24]. These higher-level representations are mostly trained in an unsupervised way, following the trend of so-called Deep Learning [3, 14, 18], but can be fine-tuned using the supervised detection objective.

These learned representations capture statistical dependencies in the co-occurrence of detections the object detectors from [24]. In fact, one can see in Table 4 plausible contexts of joint appearance of several objects learned by the CAE. These detectors, which can be quite imperfect when seen as actual detectors, contain a lot of information when combined altogether. However, the uncertainty of detectors makes it hard to combine using external semantic sources such as WordNet. As reported in Table 6, we observe a slight improvement (+0.5 %) using our Re-Ranking strategy and lesk words' similarities. The extraction of those *context semantics* with unsupervised feature-learning algorithms has empirically shown better performances but these semantics are inherent to the detectors outputs and can not be easily combined with any known predefined semantic system such as the one defined in WordNet.

In particular, we find that Contractive Auto-Encoder [33, 34] can substantially improve performance on top of *pose* PCAs as a way to extract non-linear dependencies between these lower-level OB detectors (especially when fine-tuned). They also improve greatly upon the use of the detectors as inputs to an SVM or a logistic regression (which were, with structured regularization, the original methods used by OB).

This trained post-processing allows us to reach the state-of-the-art on MIT Indoor and UIUC (85.13 % against 85.30 % obtained by LScSPM [12]) while being competitive on 15-scenes (86.44 % also versus 89.70 % LScSPM). On these last two datasets, we reach the best performance for methods only relying on object/part detectors. Compared to other kinds of methods, we are limited by the accuracy of those detectors (only trained on HOG features), whereas competitive methods can make use of other descriptors such as SIFT [12], known to achieve excellent performance in image recognition.

Besides its good accuracies, it is worth noting that the feature representation obtained by the *pose* PCA + CAE is also very compact, allowing a 97 % reduction compared to the original data (see Table 3). Handling a dense input of dimension 44,604 is not a common thing. By providing this compact representation, we think that researchers will be able to use the rich information provided by OB in the same way they use low-level image descriptors such as SIFT.

As future work, we are planning other ways of combining OB features e.g. considering the output of all detectors at a given scale and position and combine them afterwards in a hierarchical manner. This would be a kind of dual view of the OB features. Other plausible departures could take into account the topology (e.g. spatial structure) of the pattern of detections, rather than treat the response at each location

and scale as an attribute and the set of attributes as unordered. This could be done in the same spirit as in Convolutional Networks [22], aggregating the responses for various objects detectors/locations/scales in a way that takes explicitly into account the object category, location and scale of each response, similarly to the way filter outputs at neighboring locations are pooled in each layer of a Convolutional Network.

Acknowledgments We would like to thank Gloria Zen for her helpful comments. This work was supported by NSERC, CIFAR, the Canada Research Chairs, Compute Canada and by the French ANR Project ASAP ANR-09-EMER-001. Codes for the experiments have been implemented using Theano [4] Machine Learning library.

References

1. Baldi, P., Hornik, K.: Neural networks and principal component analysis: learning from examples without local minima. Neural Netw. **2**, 53–58 (1989)
2. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. Adv. Neural Inf. Proc. Sys. **19**, 153–160 (2007)
3. Bengio, Y.: Learning deep architectures for AI. Found. Trends Mach. Learn. **2**(1), 1–127 (2009). Also published as a book. Now Publishers, 2009
4. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a CPU and GPU math expression compiler. In: Proceedings of the Python for Scientific Computing Conference (SciPy), June 2010. Oral Presentation
5. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via plsa. In: In Proceedings of the ECCV, pp. 517–530 (2006)
6. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR09 (2009)
7. Espinace, P., Kollar, T., Soto, A., Roy, N.: Indoor scene recognition through object detection. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Anchorage, AK (2010)
8. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1785 (2009)
10. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)—Volume 2—Volume 02, CVPR'05, pp. 524–531. IEEE Computer Society (2005)
11. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discrimitatively trained, multiscale, deformable part model. In: CVPR (2008)
12. Gao, S., Tsang, I., Chia, L., Zhao, P.: Local features are not lonely laplacian sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
13. Goodfellow, I., Le, Q., Saxe, A., Ng, A.: Measuring invariances in deep networks. In: NIPS'09, pp. 646–654 (2009)
14. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**, 1527–1554 (2006)
15. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**, 177–196 (2001)
16. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. SIGGRAPH **24**(3), 577584 (2005)

17. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**, 417–441, 498–520 (1933)
18. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition?. In: Proceedings of the International Conference on Computer Vision (ICCV'09), pp. 2146–2153. IEEE (2009)
19. Kavukcuoglu, K., Ranzato, M., Fergus, R., LeCun, Y.: Learning invariant features through topographic filter maps. In: Proceedings of the CVPR'09, pp. 1605–1612. IEEE (2009)
20. Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P.: Exploring strategies for training deep neural networks. JMLR **10**, 1–40 (2009)
21. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
22. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient-based learning. In: Shape, Contour and Grouping in Computer Vision, pp. 319–345. Springer (1999)
23. Li, L.-J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: ICCV (2007)
24. Li-Jia Li, E.P.X., Su, H., Fei-Fei, L.: Object bank: a high-level image representation for scene classification and semantic feature sparsification. In: Proceedings of the Neural Information Processing Systems (NIPS) (2010)
25. Li-Jia Li, Y.L., Su, H., Fei-Fei, L.: Objects as attributes for scene classification. In: European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes, Crete, Greece, September 2010
26. Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I., Lavoie, E., Muller, X., Desjardins, G., Warde-Farley, D., Vincent, P., Courville, A., Bergstra, J.: Unsupervised and transfer learning challenge: a deep learning approach. In: Guyon I., Dror, G., Lemaire, V., Taylor, G., Silver, D. (Eds.) JMLR W& CP: Proceedings of the Unsupervised and Transfer Learning challenge and workshop, vol. 27, pp. 97–110 (2012)
27. Miller, G.A.: WordNet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
28. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. In: Visual Perception, Progress in Brain Research, vol. 155 (2006)
29. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV (2011)
30. Pearson, K.: On lines and planes of closest fit to systems of points in space. Phil. Mag. **2**(6), 559–572 (1901)
31. Quattoni, A., Torralba, A., Recognizing indoor scenes. In: CVPR (2009)
32. Ranzato, M., Poultney, C., Chopra, S., LeCun, Y.: Efficient learning of sparse representations with an energy-based model. In: NIPS'06 (2007)
33. Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., Glorot, X.: Higher order contractive auto-encoder. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) (2011)
34. Rifai, S., Vincent, P., Muller, X., Glorot, X., Bengio, Y.: Contracting auto-encoders: explicit invariance during feature extraction. In: Proceedings of the Twenty-eight International Conference on Machine Learning (ICML'11), June 2011
35. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vision **77**, 157–173 (2008)
36. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
37. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1349–1380 (2000)
38. Torralba, A.: Contextual priming for object detection. Int. J. Comput. Vis. **53**(2), 169–191 (2003)

39. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and composing robust features with denoising autoencoders. In: Cohen W.W., McCallum A., Roweis, S.T. (eds.) ICML'08, pp. 1096–1103. ACM (2008)
40. Vogel, J., Schiele, B.: Natural scene retrieval based on a semantic modeling step. In: Proceeedings of the International Conference on Image and Video Retrieval CIVR 2004, Dublin, Ireland, LNCS, vol. 3115, pp. 7 (2004)
41. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492. IEEE, June 2010