Constrained Local Enhancement of Semantic Features by Content-Based Sparsity

Youssef Tamaazousti[1 2] Hervé Le Borgne[1] Adrian Popescu[1]
[1]CEA, LIST, Laboratory of Vision and Content Engineering, F-91191 Gif-sur-Yvette, France
[2]University of Paris-Saclay, MICS, 92295 Châtenay-Malabry, France
{youssef.tamaazousti,herve.le-borgne,adrian.popescu}@cea.fr

ABSTRACT

Semantic features represent images by the outputs of a set of visual concept classifiers and have shown interesting performances in image classification and retrieval. All classifier outputs are usually exploited but it was recently shown that feature sparsification improves both performance and scalability. However, existing approaches consider a fixed sparsity level which disregards the actual content of individual images. In this paper, we propose a method to determine automatically a level of sparsity for the semantic features that is adapted to each image content. This method takes into account the amount of information contained by the image through a modeling of the semantic feature entropy and the confidence of individual dimensions of the feature. We also investigate the use of local regions of the image to further improve the quality of semantic features. Experimental validation is conducted on three benchmarks (Pascal VOC 2007, VOC 2012 and MIT Indoor) for image classification and two of them for image retrieval. Our method obtains competitive results on image classification and achieves state-of-the-art performances on image retrieval.

Keywords

Image-Retrieval, Image-Classification, Semantic-Features, Concept-Sparsification

1. INTRODUCTION

The problem of image recognition and retrieval in image databases is a topic of high interest in the vision community [1, 2, 14, 23, 25]. In parallel to the mainstream "bottom-up" approach based on convolutional neural networks (CNN) [2], several works adopted a "top-down" scheme to design semantically grounded image features, that we name *semantic features* in the following. Given the availability of large-scale image datasets, [14, 25] argued that an image representation based on a bench of object detectors is a promising way to handle natural images. These object detectors are more generally considered as the outputs of base classifiers. Such approaches offer a rich, high level description of images that is close to the human understanding. Moreover, they can easily integrate the advances of the "bottom-up" works that propose better

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912009

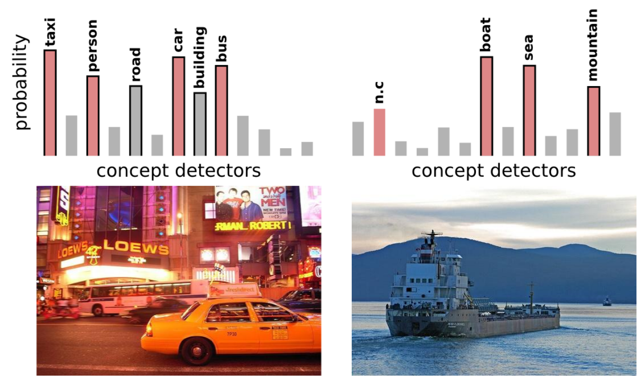

Figure 1: Illustration of the "Content-Based Sparsity" (CBS) method that adapts the sparsification of semantic features to the actual content of the image. Two images with different semantic profiles are presented, i.e. large and low number of detected concepts to the left and right, respectively. A fixed sparsity scheme would select the concepts illustrated in red and it would miss some useful concepts in the left image and would select noisy concepts (n.c.) in the right image. In contrast, CBS adapts the number of selected concepts (surrounded in black) and keeps a larger number in the left image and a lower one in the right image. Best viewed in color.

mid-level features in order to improve the base classifiers. Even more importantly, semantic features are scalable in terms of number of classes recognisable in order to cope with a wide variety of content. In image retrieval, they can be represented with inverted index implementation that is particularly efficient for large-scale databases. In classification, the compactness of the features insures computational efficiency, in particular at testing time. The classical formulation of semantic features exploits all classifier outputs [25] but it was recently shown that feature sparsification can be beneficial both in terms of scalability and performance [9].

The contribution of this paper is two-fold. First, we propose a method to determine automatically an appropriate level of semantic feature sparsity that accounts for the amount of information contained by the image that is modelled with the Shannon entropy and the confidence of the individual prediction of semantic feature dimensions. We study the importance of both components for an appropriate modelling of image content. Above the particular method proposed, such an adaptation of the level of sparsity has never been proposed to date in the domain of semantic features.

Second, inspired by Object Bank scheme [14], we also deal with the problem of the content-locality aspect in images. In Object Bank, the spatial location of the objects is encoded with a three level pyramid computed at multiple scales of the image. However, the final feature in [14] results from the concatenation of the features in all grids. While concatenation is not problematic in Object Bank since it only includes approximately 200 detectors, it has significant negative effect on scalability when exploiting tens of thousand detectors [1, 9]. As an alternative for locality modelling, we apply a pooling of local regions that does not change the total size of the feature and its sparsity level. The amount of useful information varies from one region to another and this variability is accounted through a max-pooling scheme that selects the top score of each individual concept and merges these individual scores into a single vector before sparsification. The proposed locality modelling scheme is named "Constrained Local Enhancement" (CLE) and built on top of the first contribution. The use of max-pooling and spatial pyramid is not novel in itself since it has been used in the domain of bag-features [2] and CNNs [12], but nevertheless its use is novel for semantic features, and we show it makes particular sense when it is combined with the first proposed contribution.

We validate our work on three publicly available benchmarks, focusing on a scene classification task with the MIT Indoor benchmark and on multi-class object classification task with PascalVOC 2007 and Pascal VOC 2012. Result shows that the proposed approach CLE obtains competitive results compared to the best approaches in the literature and achieves state-of-the-art performances compared to semantic-based approaches. We also validate our approach on image retrieval over two datasets, Pascal VOC 2007 and MIT Indoor 67, when CLE achieves state-of-the-art performances.

2. RELATED WORK

The current trend in image classification and retrieval is to exploit mid-level features obtained with deep convolutional neural networks, such as Overfeat [21], Caffe [13] and VGG-Net [24]. Building on such features, we focus on "semantic features" that: (i) include a rich representation of images, (ii) provide a humanly understandable description of content, and (iii) are more flexible since concepts are learned independently from one another. The *semantic feature* approach has been introduced in [14, 25] with a limited number of concepts, with several proposals to determine the level of sparsity for these features. For instance, [14] managed this sparsity aspect at learning time through the regularisation of logistic regression.[25] adopted a more direct scheme by retaining only a portion of the most probable base classifier outputs. In practice they selected $1,500$ non-zero dimensions among $2,659$ possible visual concepts. However, recent works such as [9] exhibits very good performances in image retrieval by retaining a small and fixed number of classifier outputs (*i.e.* less than 100 dimensions) for all images. To the best of our knowledge, our work is the first that proposes to adapt the level of sparsity to the content of each image.

The number of base classifiers used was limited and hand-picked in the first works [14, 25] but [1] learned a larger number of visual concepts to provide better coverage of semantic features. However, their approach is based on 13 different features and then "lifting-up" each one to approximate a non-linear kernel, resulting in a significant computational cost. In order to gain in efficiency, they showed that a binarization of their $15,458$-dimensional vector resulting from a hierarchy of $8,000$ synsets still leads to good results, though inferior to those obtained with the full feature. Recently, [9] exploited $17,000$ concepts from ImageNet or $30,000$ concepts from Flickr Groups and obtained interesting results in image retrieval using simpler linear classifiers learned over mid-level CNN features. Relating to the choice of the visual concepts of reference, we adopted a scheme similar to [9], that has the advantage to be efficient and easy to reproduce.

Regarding semantic feature locality, Object Bank [14] accounts for the spatial location of the objects using a three level pyramid computed at multiple scales of the image to reduce scale variance. As we mentioned, this concatenation based approach scales poorly when using a large number of base classifiers. However, the contribution of our paper does not relate to obtain the exact localisation of an object but to include such a possible locally present object into a global signature. Regarding the modelling of the local information, our contribution is inspired by the spatial pyramid pooling (SPP) [12]. However, in this last paper, SPP is applied at a lower level (that of the mid-level features) and requires a much larger number of regions than in our case. In our case, we show that when it is applied at a higher level (semantic feature) and it is combined with the CBS that adapt the number of active concept to the actual content of each region, SPP makes sense with a small number of regions. This latter is specifically due to the ability of our approach to consider informative regions and neglect confusing ones.

3. IMPROVED SEMANTIC FEATURES

We first present the formalism of existing semantic features [1, 9, 25]. Then, we detail the "Content-Based Sparsity" (CBS) scheme that automatically adapts the level of sparsity to the content of the image. Finally, we introduce "Constrained Local Enhancement" (CLE), the main contribution of this paper which applies CBS to image regions to integrate only informative local regions.

An image I is first described by a global feature $\mathbf{x}$. Then, a *raw* semantic feature is defined as a C-dimensional vector $\zeta^{raw}(\mathbf{x}) = [\phi_1^{raw}(\mathbf{x}), \ldots, \phi_C^{raw}(\mathbf{x})]$ where each dimension $\phi_c(\mathbf{x})$ is the output of a binary classifier evaluated on $\mathbf{x}$, further normalized by a sigmoid function, such that $0 < \phi_c(\mathbf{x}) < 1$. First works in the domain [1, 25] used the LP-β kernel combiner of [8] on 13 features. These approaches had an important computational complexity, but [9] showed that linear classifiers could be used under the constraint of (i) using a significantly larger C (ii) using more powerful mid-level features $\mathbf{x}$. Following [9], the semantic feature is learned on top of ImageNet [3] concepts that have at least 100 associated images, results in $17,462$ dimensions. To study its robustness for different mid-level features that are used to learn object classifiers, it is implemented on top of Overfeat, Caffe and VGG-Net architectures and results are discussed in Section 4.4. A diversified negative class is used when learning classifiers. This class is composed of images that are pooled from ImageNet concepts that are not included in the semantic feature. Its content is sorted in such a way a maximum of concepts are used when learning base classifiers. We empirically determined that best results are obtained for a ratio of $1/100$ between positive and negative samples and this ratio is used in experiments. Finally, we chose to learn each classification model $\phi_c(.)$ with L_2-regularized linear SVM.

Given a *raw* semantic feature, it has been shown that the performances in retrieval and classification could be almost the same when the descriptor is quantized on a limited number of bits [25, 1], or even improved when one sparsifies the feature by retaining only the d largest values and set the other to zero [9]. Formally, the signature with a "fixed" level d of sparsification is expressed as:

$$\phi_c^d(\mathbf{x}) = \begin{cases} \phi_c^{raw}(\mathbf{x}) & \text{if } \phi_c^{raw}(\mathbf{x}) \in \mathcal{H}_d(\zeta^{raw}(\mathbf{x})) \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{H}_d(A)$ is the subset of the d largest values of the set $A \subset \mathbb{R}$. The semantic feature with a fixed level of sparsification d is then

$\zeta^d(\mathbf{x}) = [\phi_1^d(\mathbf{x}), \ldots, \phi_C^d(\mathbf{x})]$. Equation (1) correspond to the "fast encoding" of locality-constrained coding [26], an efficient method of sparse coding used in the bag-of-visual-word framework. It introduces a locality constraint to the feature representation that leverages the manifold geometry induced by the set of visual concepts, locally homeomorphic to an Euclidean space [26], and improves the recognition performances in practice [9]. In addition, such a sparsification leads to a much smaller memory footprint since, if a float and an index are both coded on 4 bytes, then ϕ_c^{raw} occupies $4 \times C$ bytes while ϕ_c^d only $8 \times d$ bytes (with $d \ll C$).

3.1 Content-based Sparsity for Semantic Features

We make two hypotheses that motivates our proposal.First, the level of sparsity of a semantic feature should be adapted to the number of objects/concepts contained in the image. As illustrated in Figure 1, it seems desirable that an image containing lots of object (Fig. 1, left image) has a semantic feature with more non-zero dimensions than that of an image with few objects (Fig. 1, right image). Retaining a fixed number of concepts as it has been proposed to date leads to an incomplete description for the first image and to an over-complete one for the second one.

Second, it is equally important that the number of non-zero dimensions to retain also depends on the confidence that can be placed into the quality of visual concept detections. For instance, if the confidence of concept detections is low (bottom of Figure 2), only a small number of concepts should be retained. On the contrary, if confidence is high (top-right of Figure 2), a larger number of concepts would probably be useful in the image representation. In [25] such a selection is proposed but is based on a cross-validation error of the ϕ_c and is thus the same of all images. As well, [1] proposed a simple thresholding on the values of $\phi_c(\mathbf{x})$ that is not adapted to each image. To our kowledge, our proposal is the first to propose an adaptation to *each* image.

The combination of the two hypotheses leads to four prototypical semantic feature profiles that are depicted in Figure 2. When few visual concepts are present in the image (see left of Figure 2), the profiles include a small number of dominating values (values with high confidence) and profiles are naturally sparse. On the contrary, the presence of a large number of visual concepts (see Figure 2, right schemes) leads to a flatter profile, typical of a distributed activity. In addition, the value of the largest concept (thus on the left of the profiles in Figure 2) indicates the confidence one can place into the quality of the visual concepts detected. If the two hypotheses above are combined, we should retain a large number of concepts when the profile indicates both a confident detection and many dominant concepts (top-right of Figure 2), while we should keep few concepts for the other three cases.

Let $\zeta^{raw}(\mathbf{x}) = [\phi_1^{raw}(\mathbf{x}), \ldots, \phi_C^{raw}(\mathbf{x})]$ be the raw semantic feature. Equation (1) proposes a fixed sparsity level that does account for the two hypotheses that support an appropriate modeling of individual images. We propose to incorporate these hypotheses and thus adapt sparsity to image content using:

$$\phi_c^{CBS}(\mathbf{x}) = \begin{cases} \phi_c^{raw}(\mathbf{x}) & \text{if } \phi_c^{raw}(\mathbf{x}) \geq \Gamma(I) \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\Gamma(.)$ is a threshold that reflects "semantic feature profile" of each image, that is to say, the way its values decrease, when they are sorted in decreasing order.

Each dimension of the raw semantic signature $\phi_i^{raw}(\mathbf{x})$ can be seen as a piece of information related to the content of the image. We propose to consider the raw semantic signature of an image as a source of information on its semantic content and to model the

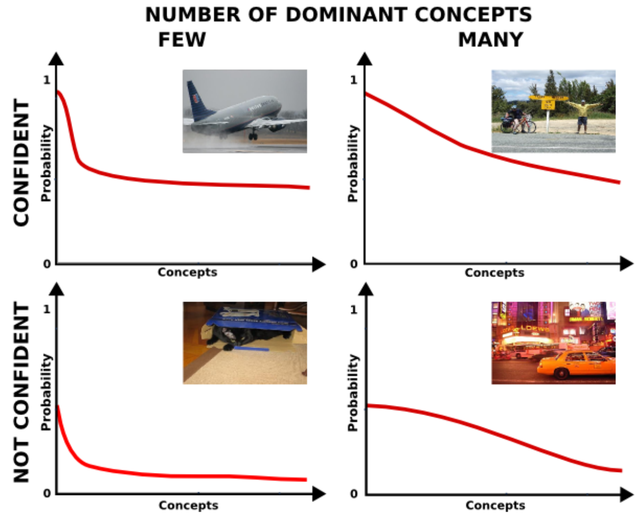

Figure 2: Four configurations of semantic feature profiles. Each configuration illustrates the raw semantic signature with $\phi_i^{raw}(\mathbf{x})$ sorted by decreasing values. The top graphics illustrate high confidence detections (well recognizable objects in the images) and the bottom ones low confidence detections (incomplete objects in the image). The left graphics correspond to the schemes containing few dominant visual concepts (few object in the image), while the right ones, depict schemes with a lot of dominant visual concepts (several objects in the image).

quantity of information it conveys using the Shannon entropy. This choice is formally supported by the fact that the entropy is defined as the average amount of information generated by a source. If appropriately normalised to be considered a random variable, ζ^{raw} can be considered as such a source of information. Formally, for an image I from which a mid-level feature $\mathbf{x}$ was extracted:

$$H(I, \mathbf{x}) = \sum_{i=1}^{C} \frac{\phi_i^{raw}(\mathbf{x})}{\sum_{j=1}^{C} \phi_j^{raw}(\mathbf{x})} \log_2 \left(\frac{\sum_{j=1}^{C} \phi_j^{raw}(\mathbf{x})}{\phi_i^{raw}(\mathbf{x})} \right) \tag{3}$$

The value $\frac{\phi_i^{raw}(\mathbf{x})}{\sum_{j=1}^{C} \phi_j^{raw}(\mathbf{x})}$ is the probability to get concept i into the image. With such a normalization, the semantic feature is the probability mass function of a discrete random variable whose value is subject to variations due to the presence of object/concepts into the image. Hence, equation 3 computes Shannon entropy of this source of information. It results into a lower entropy when profiles are distributed and higher for the profiles having a small number of dominant concepts. The threshold $\Gamma(.)$ should therefore be a decreasing function with respect to $H(I, \mathbf{x})$.

Equation 3 reflects the number of dominant concepts but does not take into account the absolute confidence of their detection. For this, we propose to consider the value of the largest semantic dimension of the profile, noted $\phi_{max}(\mathbf{x})$ that is normalized in $[0, 1]$, as the proportion of concepts to retain among the C available. For instance, for an almost flat profile, our proposal means that if the output classifiers have a confidence around 0.5 we decide to retain half of them, while for a confidence around 0.75 we retain the three quarters of them. In accordance to our hypothesis, the more confident detections are, the more dimensions will be retained. Hence, the threshold $\Gamma(.)$ should therefore be an increasing function with respect to $\phi_{max}(\mathbf{x}) \times C$.

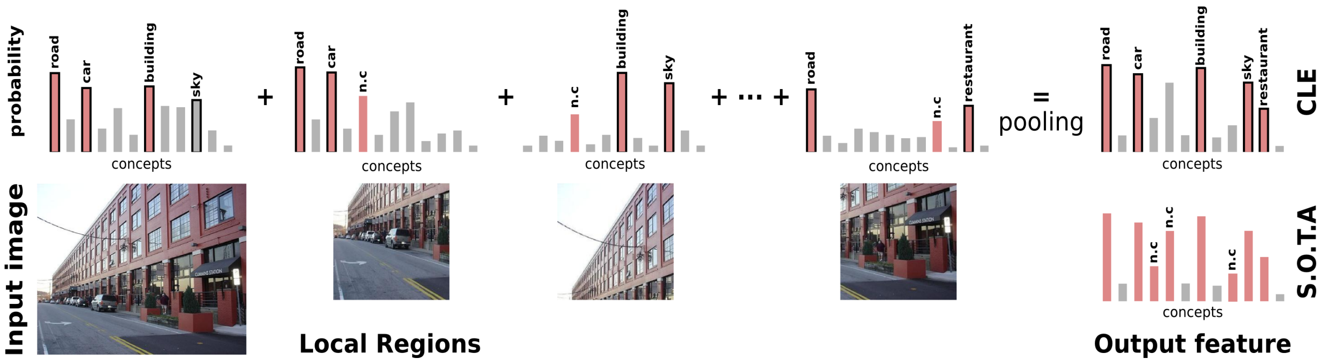

Figure 3: Illustration of our "Constrained Local Enhancement" (CLE) method over semantic features. Concepts illustrated in red are selected by the fixed sparsity, while concepts surrounded in black are selected by our "Content-Based Sparsity" (CBS). The use of fixed sparsity on local regions results in a final representation (bottom) that contains a lot of noisy concepts (n.c.). On the contrary, CBS selects only relevant concepts per region and results in a relevant output feature (top). Using max-pooling, selected concepts from the global image representation are updated with higher scores from image regions. Equally important, new relevant concepts with strong activations in image regions appear in the final representation. Best view in color.

Combining the two modeling of our hypotheses, we finally propose to estimate the threshold as:

$$\Gamma(I) = \alpha \times \frac{\phi_{max}(\mathbf{x}) \times C}{H(I, \mathbf{x})} \tag{4}$$

where the denominator $H(I, \mathbf{x})$ is the entropy computed according to equation (3), $\alpha \in [0, 1]$ is a normalizing parameter that can be set by cross-validation, C is the total number of visual concepts that are considered in the semantic signature and $\phi_{max}(\mathbf{x})$ the largest value of the semantic feature, named "confidence parameter".

3.2 Constrained Local Enhancement

The mid-level CNN features used to produce the semantic features are not scale-independent [19] and small objects have smaller detection scores if only the full image is exploited. In Figure 3, we illustrate the importance of using local regions with the examples of *car*, whose score is taken from a local region and *restaurant*, which does not appear at all in the representation of the full image. We enhance image semantic features by adding local image information. Let's consider an image I from which we identify a set of regions $\{\mathcal{R}_i, i \in [1, N]\}$. A mid-level feature $\mathbf{x}_i$ is extracted from each region and the corresponding semantic feature $\zeta_i^{d_i}(\mathbf{x}_i)$ is computed with a level of sparsity d_i. Then, the final feature of the image results from the pooling of all these local semantic features:

$$\zeta_{loc}(I) = \underset{i=1\ldots N}{\mathcal{P}}(\zeta_i^{d_i}(\mathbf{x}_i)), \tag{5}$$

where $\mathcal{P}$ is a pooling operator such as sum, $average$ or max that operates on individual components of semantic features of image regions. The pooling procedure is well established by biophysical evidence in visual cortex [22] and is empirically justified by many algorithms applied to image categorisation.

One advantage of the method is that, unlike concatenation-based region modeling schemes such as spatial pyramid matching (SPM) [10, 16], the resulting feature has the same size as the original semantic feature computed on the full image. Moreover, it contains information extracted at a local level that represents more faithfully the content of the image.

Naturally, the performance of the local enhancement of the feature depends on the choice of initial regions $\mathcal{R}_i$. Here we propose a simple strategy in order to demonstrate the efficiency of the principle and adopt a scheme inspired by SPP [12]. The region $\mathcal{R}_0$ is the full image and the following $\mathcal{R}_{i>0}$ are extracted according to regular grid at smaller scale. Unlike SPM, we use overlapping rectangular regions to reduce the risk of cutting the objects represented in the image.

A semantic feature can be enhanced locally with a fixed-sized sparsity (d_i fixed $\forall i \in [1, N]$), but it results into the assignment of the same number of visual concepts to all local regions. Such an approach is suboptimal since some regions could contain a lot of information while others may contain much less. Hence, we hypothesize that we should assign automatically a weight to each region, in order to determine which ones should be most considered. As illustrated in Figure 3, our first contribution, "Content-Based Sparsity" (CBS) method, has the ability to assign a sparsity level to regions based on their amount of information. Finally, when the level of sparsity d_i used in equation 5 is determined by the CBS method (equation 4) on all regions before the pooling, we obtain our main contribution, namely "Constrained Local Enhancement" (CLE). Formally, the final feature is computed through

$$\zeta_{loc}(I) = \underset{i=1\ldots N}{\mathcal{P}}(\zeta_i^{CBS}(\mathbf{x}_i)), \tag{6}$$

where $\mathcal{P}$ is a pooling operator such as sum, $average$ or max that operates on individual components of semantic features of image regions. The sparsification of each region is computed by our "Content-Based Sparsity" (CBS) defined by Eq. (2) and (4).

4. EXPERIMENTS

The effectiveness of our approach is tested in image classification (Section 4.2) and image retrieval (Section 4.3) tasks. To facilitate reproducibility and comparability, evaluation is done with publicly available datasets and using standard experimental protocols. Image classification evaluation is conducted on Pascal VOC 2007 [7], Pascal VOC 2012 [6] (object recognition) and MIT Indoor 67 [18] (scene recognition). For image retrieval, we evaluate our approach on Pascal VOC 2007 and MIT Indoor 67. [1]

The proposed "Constrained Local Enhancement" method can be applied to a semantic feature built on top of any mid-level feature.

[1]Pascal VOC 2012 is excluded due to the unavailability of an evaluation protocol and ground truth for image retrieval.

However, based on the experiments reported in Section 4.4, the quality of the semantic feature will directly depend on that of the mid-level feature used. We thus created semantic features on top of a competitive mid-level CNN feature released in the literature, namely VGG-Net [24]. For our study, fine tuning of the CNN may result into an improvement of the results at the cost of significant computational cost and the possible use of additive data. Such a specific optimization of the CNN has not been considered in our experiment, to insure their reproducibility with original and available CNN models.

4.1 Baseline Methods

Our work is focused on semantic features and it is important to compare the performance of the proposed CLE method to several state-of-the-art semantic-based approaches. We also compare it to a very competitive CNN feature.

- **VGG-Net** [24], is extracted from a fully-connected layer (fc7, 16^{th} layer) of a CNN architecture (D) learned on ILSVRC 2012 dataset [20] that contains 1.2 million images of 1,000 classes. The resulting vector has $4,096$ dimensions. The 18^{th} and last layer (fc8), of size $1,000$, can be seen as a semantic feature build on top of the fc7, which the base classifiers are the final outputs of the CNN.

- **Semfeat** [9] is built on top of a mid-level CNN feature using the Caffe reference model [13]. To ensure a fair comparison with our method, we rebuild Semfeat using VGG-Net as basic feature. The same 17,462 ImageNet concepts from [9], represented by at least 100 images, are modeled here. A fixed sparsification is used to replicate their methodology.

- **Classemes+**, is our own implementation of **Classemes** [25]. Again, for a fair comparison with other methods, it is built on top of a VGG-Net feature with exactly the same concepts as Semfeat. Following [25], no sparsification is considered and **Classemes+** thus corresponds to ζ^{raw}.

- **Meta-Class** [1] is the output of 15,232 classifiers. It is based on a concatenation of five low-level features combined with a spatial pyramid histogram with 13 pyramid levels. Since the number of concepts is rather similar to other methods and the code is available, we use it as it is released. [2]

4.2 Image Classification Experiments

We report and analyse the experimental results on image classification task using the Pascal VOC 2007, Pascal VOC 2012 and MIT Indoor 67 datasets.

4.2.1 *Object Classification*

The Pascal VOC 2007 object classification task [7] is run on a dataset that contains $9,963$ images. Each image is labelled with one or more categories from a total of 20. We used the pre-defined split of $5,011$ images for training and $4,952$ for testing, with the publicly available evaluation tools and ground truth. The Pascal VOC 2012 benchmark [6] is similar to VOC 2007 but its number of images is larger: $22,531$ images are split into training ($5,717$ images), validation ($5,823$ images) and test data ($10,991$ images). A server is available to estimate the performances of an algorithm on the test dataset, with a limited number of submission allowed. Hence, we measured the performance on the official testing dataset.

[2] http://vlg.cs.dartmouth.edu/projects/metaclass/metaclass/Home.html

For both datasets, we learn each class by a one-vs-all SVM classifier and we use mean Average Precision (mAP) to evaluate performance. The cost parameter of the SVM classifier and the α parameter from equation (4) are optimised through cross-validation on the training dataset, using the usual train/val split. In addition to the baselines presented in Subsection 4.1, we also report the best existing state-of-the-art results in the literature.

Classification results for Pascal VOC 2007 and 2012 are presented in Table 1. Our method leads to significant better results than previous approaches based on semantic features [25, 1, 9], as well as those obtained from the mid-level feature it is based on (VGG-Net [24]). Note that we get better results than VGG-16 (our method is built on top of this mid-level features) but [24] report results above ours with VGG-19, a multi-scale VGG-Net feature of 19 layers. However, this is obtained when descriptors are aggregated over five scales and they note that their performances is much below when only three scales are used (although they do not give the corresponding score). On our side, we obtain 88.2 with only two scales and a VGG-Net features of 16 layers only. The improvement due to our method with regards to the mid-level feature used is discussed in Section 4.4. In the same vein, Wei *et al.* [27] achieves 2% better mAP on VOC12 when their method is trained on an extended 2000-classes ILSVRC dataset but is 5% of mAP below our method when they use a comparable dataset as ours to train their CNN. Note that our proposal is always better on VOC07 (from 3% to 7% mAP), even in comparison to the method trained with the extended dataset.

Method		**VOC 2007** mAP (in%)	**VOC 2012** mAP (in%)
CNN	Oquab *et al.* [19]	77.7	n.a.
	Chatfield *et al.* [2]	82.42	83.2
	Wei *et al.* [27]	81.5(85.2**)	81.7(**90.3****)
	VGG-Net (fc7) [24]	86.1(**89.7***)	84.5 (89.3*)
Semantic	VGG-Net (fc8) [24]	77.4	77.2
	Meta-Class [1]	48.4 (53.2†)	49.3
	Classemes+ [25]	82.4	81.7
	Semfeat [9]	82.8	81.7
	CLE (ours)	**88.2**	**86.6**

Table 1: Comparison with the state of the art on Pascal VOC 2007, Pascal VOC 2012. Our model is denoted as "CLE". Results marked with * where achieved using multi-scale scheme with a CNN architecture (E) of 19 layers. Results marked with ** were achieved using a CNN pre-trained on the extended ILSVRC dataset (2000 classes). Results marked with † are reported scores in the original paper.

4.2.2 *Scene Classification*

The MIT Indoor database [18] is a scene recognition benchmark that consists of 67 categories of indoor places. All experiments in this database where performed using the usual training-test split, taking 80 images per class for training and 20 for testing. It results in $5,360$ images in training and $1,360$ in testing.

We learn a SVM classifier on each class (one-versus-all strategy) and cross-validate the cost parameter and the α parameter of the equation (4) by splitting the training dataset into 60 images per class for training and 20 images per class for validation. Similar to Pascal VOC experiments, the proposed CLE method is compared to related baselines but also with other state-of-the-art approaches that were evaluated MIT Indoor 67. All results are reported in the

Table 2. We report classification accuracy to evaluate the performances. Zhou *et al.* [30] report an accuracy of 56.79 on MIT Indoor 67, using a generic CNN architecture similar to Caffe. With this mid-level feature, our method gets 57.3. They also report a score of 68.24 when they fine-tune the network with 2.5 million additional images representing placs. When they combine with the original ImageNet images and remove the overlapping scene categories, they obtain an Hybrid-CNN trained on 3.5 million images from 1,183 categories that report 70.8 classification accuracy. Our CLE method based on VGG-Net obtain better results (71.6) while we use a CNN pre-trained on object images only while their deep network is adapted to recognize places. Our proposed approach also outperforms over all semantic-based approaches and even other state-of-the-arts works such as ONE+SVM [28].

	Method	**MIT Indoor 67** Classification Accuracy (in%)
	Doersch *et al.* [4]	66.9
	Oquab *et al.* [19]	69.0
	VGG-Net (fc7) [24]	65.8
	Zhou *et al.* [30]	68.2* (70.8**)
	ONE+SVM [28]	70.1
Semantic	VGG-Net (fc8) [24]	48.7
Semantic	Meta-Class [1]	35.7 (44.6†)
Semantic	Classemes+ [25]	58.9
Semantic	Semfeat [9]	61.5
Semantic	CLE (ours)	**71.6**

Table 2: Comparison with the state of the art on MIT Indoor 67. Our model is denoted as "CLE". Results marked with * where achieved using a CNN pre-trained on 2.5 million scene images. Results marked with ** where achieved using the previous pre-trained CNN with fine-tuning. Results marked with † are reported scores in the original paper.

4.3 Image Retrieval Experiments

We report and analyse the experimental results on visual retrieval over Pascal VOC 2007 and MIT Indoor 67 datasets. Pascal VOC 2012 has not been evaluated because of the absence of ground-truth on test images. For both evaluated datasets, we adopt Average Precision at top K retrieved results (AP@K) defined as in [29]:

$$AP@K = \frac{1}{min(R,K)} \sum_{j=1}^{K} \frac{R_j}{j} \times I_j, \qquad (7)$$

where R is the number of relevant images within the collection and R_j is the number of relevant images among top j search results. I_j is set to 1 if the j-th image is relevant, and 0 otherwise. We averaged AP@K over all queries to obtain mAP@K as overall evaluation metric. $K \in \{1, 2, \cdots, 100\}$ for Pascal VOC 2007 and $K \in \{1, 2, \cdots, 80\}$ for MIT Indoor 67, because the number of relevant images within the dataset is set to 80 per query. Similarity between images are computed using cosine measure.

Object retrieval evaluation is run with the Pascal VOC 2007 dataset, using the train images as collection and the test images as queries. In Figure 4, we present mAP@100 curves for CLE, Classemes+ and Semfeat, the best semantic feature baselines, and also for VGG-Net. Figure 5 reports detailed performance (AP) of individual concepts for the CLE, Semfeat VGG-Net.

Scene retrieval is evaluated with the MIT Indoor 67 dataset, using the same split as for classification, with training images considered

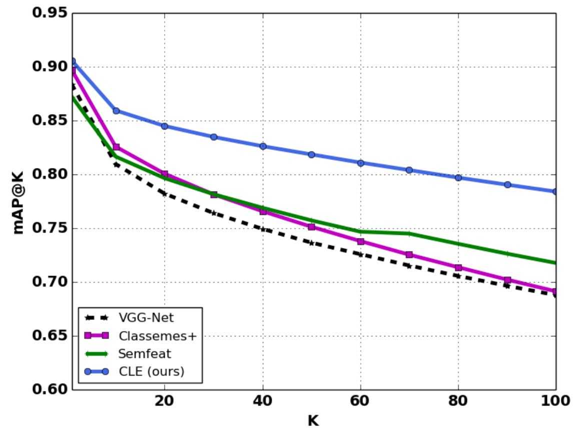

Figure 4: Overall performance (mAP@K) of visual retrieval using all images of the released test split as queries on Pascal VOC 07. We compare our CLE method, with the best semantic-based state-of-the-art approaches (Classemes+ [25] and Semfeat [9]) and the best CNN-based method (VGG-Net [24]) represented in dashed line.

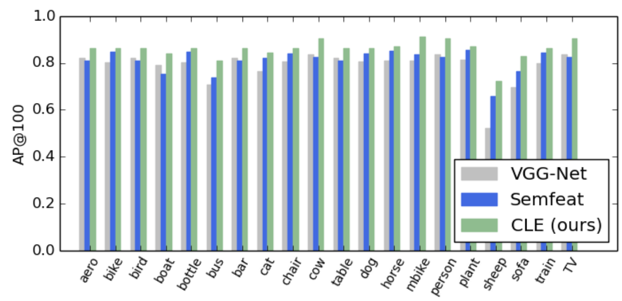

Figure 5: Detailed performance (AP@100) of individual categories for the best semantic-based (Semfeat [9]) and CNN-based (VGG-Net [24]) approaches reported in the Figure 4 for image retrieval on Pascal VOC 2007 dataset.

to be the collection and test images considered as queries. Figure 6 gives mAP@K curves for CLE, Classemes+, Semfeat and VGG-Net. For each query, there are at most 80 true positives in the ground-truth and we set K=80. Figure 7 reports detailed performance (AP) of individual concepts for the best semantic-based (Semfeat) and CNN-based (VGG-Net) approaches.

The obtained results show that the proposed CLE method outperforms all other tested methods on both of the evaluated datasets. For instance, on Pascal VOC 07, CLE improves the performance of VGG-Net, Classemes+ and Semfeat by around 10%, 8% and 7% in terms of mAP@100, respectively. CLE achieves best results over all methods (CNN and semantic-based) on all categories. On MIT Indoor 67, CLE improves the performances of VGG-Net, Classemes+ and Semfeat by around 6%, 7% and 5% in terms of mAP@20, respectively. Nevertheless, we observe on the detailed performances (AP@20) of individual categories that, for a few categories, VGG-Net performs better, this can be due to the fact that some scene categories like *pool inside*, do not contain any object but only basic shapes, while CLE has the ability to recognize ob-

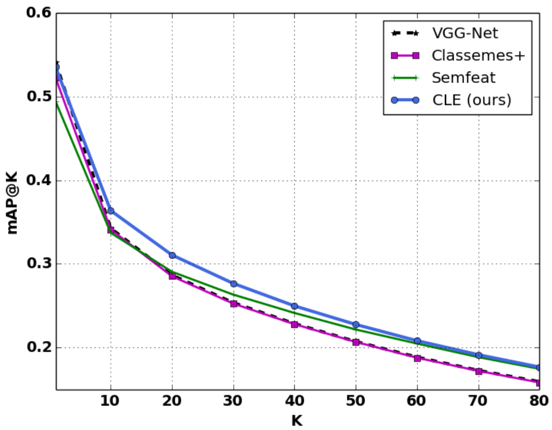

Figure 6: Overall performance (mAP@K) of visual retrieval using all images of the released test split as queries on MIT Indoor 67. We compare our CLE method, with the best semantic-based state-of-the-arts approaches (Classemes+ [25] and Semfeat [9]) and the best CNN-based method (VGG-Net [24]) represented in dashed line.

jects only. Based on that, our proposed CLE outperforms all the methods when scene categories contain objects. The above phenomenons of performance comes from an effective cooperation of local representations and a Content-Based Sparsity that keeps in the final representation only relevant concepts. On the contrary, Semfeat forces a level of sparsification and Classemes+ keeps all the concepts in the final representation, yielding in a consideration noisy concepts in the final feature.

4.4 Mid-level Feature Sensitivity

The low-level or mid-level features used to build concept detectors are a core component of semantic features. We evaluate their influence on performance by using three publicly available CNN models using the Pascal VOC 2007 dataset. Overfeat [21], Caffe [13] and VGG-Net [24] are used as basic features. Overfeat and Caffe architectures are similar since both of them are based on the original AlexNet model [15]. VGG-Net [24] uses a deeper CNN architecture, with smaller kernels and exhibits significantly better results than Overfeat and Caffe on the ImageNet challenge. In all cases, we extracted features from the last fully-connected layer, resulting in $4,096$ dimensional vectors. Regarding the resulting semantic features, note that, Classemes was initially designed with low-level features only, as well as Semfeat, was initially designed with Overfeat features only. Thus, we implemented Classemes [25], Semfeat [9] and our CLE method on top of the three mid-level features (Overfeat, Caffe and VGG-Net). For Semfeat, the sparsity parameter is fixed as formalised by equation (1). It has been cross-validated on several levels of sparsity, and we report the best results, corrseponding to $d = 50$. For our "Constrained Local Enhancement" (CLE), the parameter α of equation 4 has been cross validated on the validation data and we report the best results, corresponding to $\alpha = 0.1$.

Results are reported in Table 3 and, as expected, they are correlated to the quality of the mid-level feature used. For all the semantic methods, best performance is obtained with VGG-Net followed by Caffe and Overfeat. This suggests that, the more the mid-level features are deeper, the more the obtained results by semantic features will be better. Equally important, regardless the mid-level feature used, our CLE approach is better than other semantic-based ones. This validates the robustness of our approach and, since other mid-level features are easy to plug into the CLE pipeline, it will be easy to make it evolve in order to take advantage of progress in deep learning architecture design.

Method	**Pascal VOC 2007** mAP (in %)		
	Overfeat [21]	Caffe [13]	VGG-Net [24]
CNN only	72.0	76.3	86.1
Classemes+ [25]	72.2	72.4	82.4
Semfeat [9]	73.6	76.0	82.8
CLE (ours)	78.4	80.5	88.2

Table 3: Evaluation of effects of CNN architecture (Overfeat, Caffe and VGG-Net) used to compute semantic features (Classemes, Semfeat and Ours) on object classification over Pascal VOC 2007 dataset.

5. CONCLUSIONS

We introduce a novel method to design semantic features that integrates adaptive sparsification and information from informative local regions only. In contrast to existing works, which perform sparsification regardless of image content, we propose a scheme that considers individual image profile to do so. The informational content of images is modelled using the Shannon entropy, a theoretically grounded method, and also accounts for the confidence that can be placed in visual concept detections. Finally, modelling local regions of the image further improves the semantic features through an improved integration of localized objects.

Evaluation is carried out for classification and retrieval tasks and results show that the proposed method outperforms existing semantic features and also the mid-level CNN feature it is based upon. Interestingly, our CLE scheme is the only semantic feature whose classification performance is above that of the mid-level features, with existing semantic features all having lower performance. Equally important, the proposed approach is also competitive with optimizations of CNN features reported in the literature and even state-of-the-art on the MIT Indoor 67 benchmark. Content-based image retrieval results place CLE favourably when compared to all other tested approaches. In addition to performance, it is also important to note that CLE representations are sparse and thus more scalable than CNN features. Sparsity is particularly important for retrieval since images can be efficiently represented using inverted index structures which accelerate the search process.

The results obtained for still images are very encouraging and we will pursue the work reported here. We will investigate finer ways to model local information. In particular, interesting region detection [5] will replace the fixed regions that are currently used. A second work direction concerns video classification and retrieval that can be dealt with the CLE pipeline. Replacing local regions by frames in CLE, will bring to our proposed approach, the ability to assess the amount of information of each frame. This information can be used in future work to stop-frame removal [11] (removing less informative frames) or even concept-prototyping [17] (selecting a set of relevant frames) for web-videos.

Acknowledgments
This work is supported by the USEMP FP7 project, partially funded by the European Commission under contract number 611596.

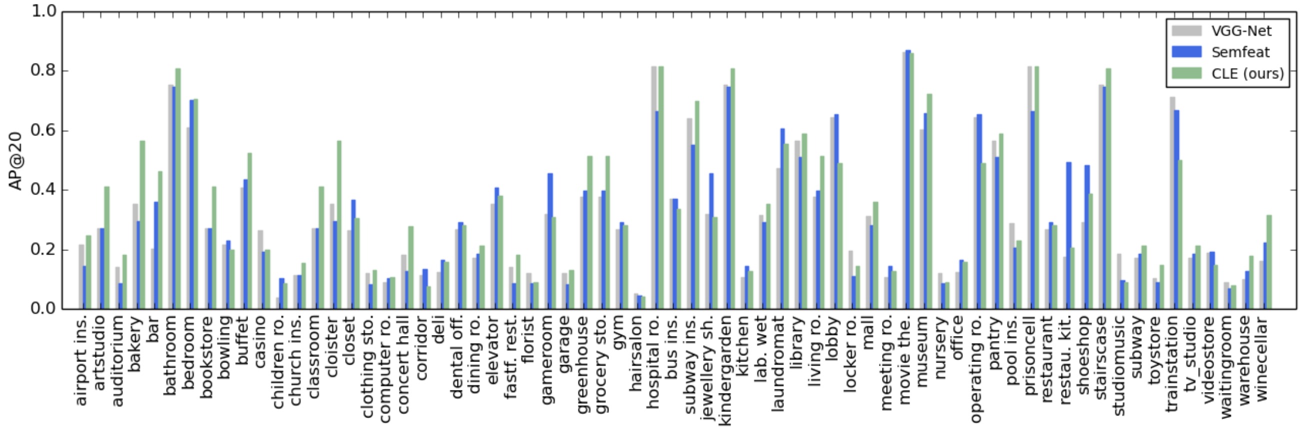

Figure 7: Detailed performance (AP@20) of individual categories for the best semantic-based (Semfeat [9]) and CNN-based (VGG-Net [24]) approaches reported in the Figure 6 for image retrieval on MIT Indoor 67 dataset. Our model is denoted as CLE.

6. REFERENCES

[1] A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. In *CVPR*, 2012.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv:1405.3531*, 2014.

[3] 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*, CVPR, 2009.

[4] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *Advances in Neural Information Processing Systems*, NIPS. 2013.

[5] P. Dollár and C. L. Zitnick. Fast edge detection using structured forests. *PAMI*, 2015.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012.

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[8] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 221–228, Sept 2009.

[9] A. L. Ginsca, A. Popescu, H. Le Borgne, N. Ballas, P. Vo, and I. Kanellos. Large-scale image mining with flickr groups. In *Multimedia Modelling*, MM, 2015.

[10] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *International Conference on Computer Vision*, ICCV, 2005.

[11] A. Habibian and C. G. Snoek. Stop-frame removal improves web video classification. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *arXiv 1406.4729*, 2014.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *International Conference on Multimedia*. ACM, 2014.

[14] L. jia Li, H. Su, L. Fei-fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*. 2010.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, NIPS, pages 1097–1105. 2012.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, CVPR, 2006.

[17] M. Mazloom, A. Habibian, D. Liu, C. G. Snoek, and S.-F. Chang. Encoding concept prototypes for video event detection and summarization. In *International Conference on Multimedia Retrieval*, ICMR, 2015.

[18] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition*, CVPR, 2009.

[19] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229*, 2013.

[22] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition*, CVPR, 2005.

[23] M. Shi, Y. Avrithis, and H. Jégou. Early burst detection for memory-efficient image retrieval. In *Computer Vision and Pattern Recognition*, CVPR, 2015.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[25] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision*, ECCV, 2010.

[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[27] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv:1406.5726*, 2014.

[28] L. Xie, Q. Tian, R. Hong, and B. Zhang. Image classification and retrieval are one. In *International Conference on Multimedia Retrieval*, ICMR, 2015.

[29] Y. Yang, H. Zhang, M. Zhang, F. Shen, and X. Li. Visual coding in a semantic hierarchy. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 59–68. ACM, 2015.

[30] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, NIPS, 2014.