Diverse Concept-Level Features for Multi-Object Classification

Youssef Tamaazousti[1 2] Hervé Le Borgne[1] Céline Hudelot[2]
[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}@cea.fr, celine.hudelot@centralesupelec.fr

ABSTRACT

We consider the problem of image classification with semantic features that are built from a set of base classifier outputs, each reflecting visual concepts. However, existing approaches consider visual concepts independently from each other whereas they are often linked together. When those relations are considered, existing models strongly rely on image low-level features, yielding in irrelevant relations when the low-level representation fails. On the contrary, the approach we propose, uses existing human knowledge, the application context itself and the human categorization mechanism to reflect complex relations between concepts. By nesting this human knowledge and the application context in the concept detection and selection processes, our final semantic feature captures the most useful information for an effective categorization. Thus, it enables to give good representation, even if some important concepts failed to be recognized. Experimental validation is conducted on three publicly available benchmarks of multi-class object classification and leads to results that outperforms comparable approaches.

Keywords

Image-Classification, Semantic-Features, Category-Level

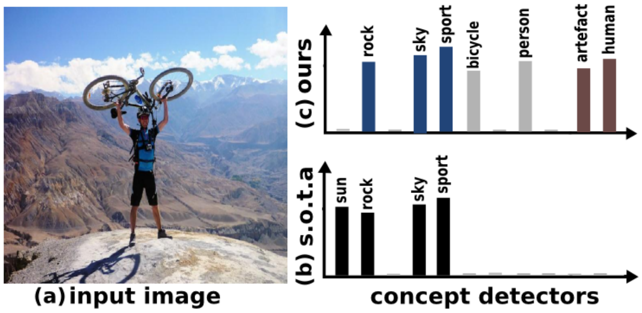

Figure 1: We propose a semantic representation that compute the concepts presence differently according to their categorical level. For an input image (a) with multiple objects, state-of-the-art semantic features (b) would output the concepts illustrated in black and miss useful concepts, such as person and bicycle. In contrast, the proposed scheme (c) capture properties of the image that are useful for categorization, *e.g. superordinate* (brown), *basic-level* (gray) and *subordinate* (blue) concepts, making the representation more relevant. Best viewed in color.

1. INTRODUCTION

The problem of object class recognition in large scale image databases is a topic of high interest in the vision community [1, 3, 14, 24, 26]. In parallel to the mainstream data-driven approach, based on convolutional neural networks (CNNs) [3, 24], several works adopted a concept-driven scheme to design semantically grounded image features, that we name *semantic features* in the following. Given the availability of large-scale image datasets, [14, 26] argued that an image representation based on a bench of object detectors is a promising way to handle natural images according to their category. 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 benefit from the advances of the data-driven works that propose better mid-level features in order to improve the base classifiers. Semantic features are also scalable in terms of number of concepts thus being able to cope with a wide variety of content.

Most semantic features in the literature [2, 10, 12, 14, 26] consider visual concepts independently from each other whereas they are often linked together by some semantic relationships (*i.e.*hyponymy, hypernymy, exclusion, etc.). An exception is the work of Bergamo and Torresani [1] that introduces "meta-classes" to address this aspect. Those meta-classes are "abstract" categories (do not really exist in the real-world) that capture common properties among many object classes. They are built using spectral clustering on low-level features of images among a set of categories. The restrictive assumption of this method, is the dependence of the meta-class learning to the visual low-level features. For instance, it leads to irrelevant meta-classes when low-level feature fails to capture the dissimilarity between different categories, making this method a "bottom-up" scheme.

The classical formulation of semantic features exploits all

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

DOI: http://dx.doi.org/10.1145/2911996.2912013

classifier outputs [1, 2, 14, 26] but it was recently shown that forcing the semantic representation to be sparse (by setting the lowest values to zero) can be beneficial both in terms of scalability and performance [10, 12]. Nevertheless, semantic features with a large set of concept detectors often contains a high number of visually similar concepts to describe the same object. For instance, the right image of Figure 2 would be predicted by a semantic feature as a *palm cockatoo*, but also a *cockatoo*, a *parrot*, a *bird*, a *vertebrate*, and so on, inducing redundant information in the final representation. As far as a human is concerned, he would categorized this image as *a bird*, an *animal* and maybe a *palm cockatoo* if the human is a bird-expert. In fact, psychologists studies such as those of Rosch [20] and Kosslyn [15] showed that a human tends to categorize an object through three categorical-levels (i) basic-level, (ii) superordinate, and (iii) subordinate. They are the most important concept types to categorize objects.

In this paper, we take into account the relations between concepts using human existing knowledge, such as semantic hierarchies (*e.g* WordNet [17]), which makes our approach a "top-down" scheme. More precisely, our main contribution consists in identifying three types of concepts into an existing hierarchy, according to their categorical level, then process them differently to design the semantic feature. It is nevertheless not easy to determine to which categorical-level a concept belongs to. Hence, we propose a method to identify the three groups in practice, for a given supervised classification problem. The proposed semantic representation is named Diverse Concept-Level feature (**D-CL**).

Compared to bottom-up approaches, an advantage of the proposed top-down scheme appears when the concept detectors fails at the subordinate level (*e.g.* the concepts *cockatoo* and *parakeet* are highly activated), which is often the case since the category is finer thus harder to identify. In that case, our descriptor at least capture basic-level and superordinate concepts (*e.g.* bird and animal), making the full representation more robust for classification problems. Moreover, the proposed feature contains only useful concepts (from the three categorical levels), which avoids redundant information that disturbs the image classification.

We validate the proposed Diverse Concept-Level representation, in a multi-object classification task through Pascal VOC 2007, Pascal VOC 2012 and Nus-Wide Object. The experiments show that the proposed approach obtains better results than seven state-of-the-art semantic features.

The remainder of this paper is organized as follows. Section 2 briefly introduces related works. Section 3 details the proposed technical approach. After showing experiments and analytic studies in Section 4, we conclude in Section 5.

2. RELATED WORKS

The current trend in image classification is to exploit mid-level features obtained with deep convolutional neural networks, such as Overfeat [23], Caffe [13] or VGG-Net [24]. Built on top of such mid-level representations, 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 each other. This semantic-based approach has been introduced by Torresani *et al.* [26] and Li *et al.* [14] with a limited number of concepts. The former used nonlinear LP-β [9] classifiers to learn each concept detector. Recently, Ginsca *et al.* [10] and Jain *et al.* [12] explored linear SVMs in semantic features and shows their effectiveness when the features are constrained to be sparse.

The feature is said *sparse* when, for a given image, only a limited number of dimension is non-zero. For instance, Li *et al.* [14] managed this sparsity aspect at learning time through the regularization of logistic regression by L_1 or L_1/L_2 [27]. Torresani *et al.* [26] did not investigate directly the sparse aspect but showed that classemes was quite robust to a 1-bit quantization. In practice, they forced negative outputs to zero thus they actually performed a sparsification. The difference with further works is that they also unified positive outputs to 1. Recent works such as [10, 12] exhibit very good performances in image retrieval, and action classification on videos, by retaining in the final feature, a small (but fixed) number of the largest classifier outputs (less than 100). In our scheme, the sparsity is a consequence of the proposed concept groups identification . Thus, in terms of sparsity, the key novelty of our work is the selection of concepts regarding their identified categorical-level, yielding to representations containing only useful concepts. Contrary to former works, our sparse representation is adapted to each image, according to its actual content, and relative to the problem of interest.

Regarding the general concepts in semantic features, Bergamo *et al.* [1] proposed "meta-classes", corresponding to "abstract" categories that captures common properties among many object classes. In our work, we also use general concepts, that captures common properties among object classes, but the key difference with their work is the building of these categories. While [1] automatically built the meta-classes using spectral clustering on low-level features of images among a set of categories, our scheme rely on a direct selection of concepts among those of WordNet [17], that have the advantage to match a semantic reality and is thus more relevant on a user point-of-view.

A last line of work deals with cognitive studies in pattern recognition [6, 16, 18, 19]. The main goal of these works is to propose a system that takes as input a set of predicted concepts for an image and outputs the corresponding basic-level concepts. In particular, the work of Deng *et al.* [6] is related to ours since they optimize the trade-off between accuracy and specificity. In other words, if concept detector fails to recognize categories at a specific level, they try to output a more general concept. As in our work, they indirectly reflect in their systems the psychological fact that even if humans tend to categorize an object at a particular level (basic-level), they are still aware of the other levels of categorization. However, the key difference between their work and ours is the method used to integrate this psychological fact as well as its purpose. In fact, our goal is to identify the most important concepts to retain in the semantic feature. In contrast, their goal is to annotate the images, and consequently identify only one concept. More precisely, we opt for an integration of the psychological fact directly in the semantic feature design, while they do it only on the test images, after the prediction of the different concepts.

3. PROPOSED APPROACH

In this section, we detail our proposed approach, a new semantic representation of an image that take into account an available human knowledge. In Section 3.1, we present our main contribution, consisting in identifying three types of concepts into an existing hierarchy (according to their

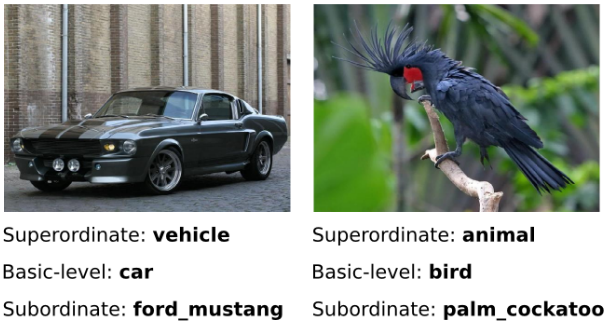
Superordinate: **vehicle**
Basic-level: **car**
Subordinate: **ford_mustang**
Superordinate: **animal**
Basic-level: **bird**
Subordinate: **palm_cockatoo**

Figure 2: Illustration of concepts that our D-CL feature would predict, for two different images. It select concepts from different categorical levels of a semantic hierarchy, *i.e.*, superordinate, basic-level and subordinate concepts.

categorical level) and then, process their concepts differently. It is nevertheless, not evident to identify these three groups, in practice. Thus, we detail in Section 3.2 how to identify them, for a given supervised classification problem.

3.1 Diverse Concept-Level Feature

A semantic feature is a F-dimensional vector $\mathcal{F}(\mathbf{x}) = [\mathcal{F}_1(\mathbf{x}), \ldots, \mathcal{F}_F(\mathbf{x})]$ extracted from an image I, itself described by a mid-level feature $\mathbf{x}$. The feature $\mathbf{x}$ could be any image descriptor such as Bag-of-Word or Fisher Kernel [11] features, but also mid-level features such as that obtained from a fully-connected layer of a convolutional neural network. Each dimension $\mathcal{F}_i(\mathbf{x})$ of the semantic feature is the output of a classifier for the concept c_i evaluated on $\mathbf{x}$.

While the concepts c_i are potentially linked together by some semantic relationships, most of works consider them independently [2, 10, 12, 14, 26]. A notable exception is the work of Bergamo and Torresani [1] that take into account relations between categories through a "bottom-up" scheme. However, their method can lead to irrelevant identification of relation when the low/mid-level features used fails to capture the dissimilarity between different categories. To cope with such a limitation, we propose to rely on existing human knowledge regarding the relations between concepts. Such a knowledge is, for instance, reflected into existing hierarchies such as WordNet [17] that organize a large set of concepts according to "is-a" relationships, that is to say by defining hyponyms and hypernyms. An advantage of our approach is to remove the dependence to the basic visual descriptor and to introduce human-based information within the process of image representation design.

All the concepts considered in semantic features, are named according existing *categories*. Once again, the *name* of a category is given according to a human judgment, and the exact choice of the word used is far from being neutral, as a large literature has shown it, both in Psychology [15, 20] and Computer Vision[6, 16, 19]. More precisely, they showed the importance to differentiate several levels of categories:

- **Basic-level concepts** are the terms at which most people tend naturally to categorize objects, usually neither the most specific nor the most general available category but the one with the most distinctive attributes of the concept.
- **Superordinate concepts** are categories placed at the top of a semantic hierarchy and thus displays a high degree of class inclusion and a high degree of generality. They include basic-level and subordinate concepts.
- **Subordinate concepts** are found at the bottom of a semantic hierarchy and display a low degree of class inclusion and generality. As hyponyms of basic-level concepts, subordinate categories are highly specific.

At the core of our proposal to design a feature representation, concepts are processed differently according to their categorical level. This asymmetrical process is based on a cognitive study proposed by [15] where they conclude that, concepts are processed differently by humans, *i.e.*, it is purely perceptual for the basic-level and subordinate concepts, while it is inferred using stored semantic information, for superordinate concepts. In our scheme, basic-level and subordinate concepts are computed through a visual process, while superordinate concepts are processed semantically using the hyponym relations between concepts into hierarchies. Figure 2 illustrates for two input images, the three types of concepts that would be retained by our scheme.

More precisely, for an input image, the probability of a basic-level or a subordinate concept is the output of a binary classifier ($\varphi_i^V(\mathbf{x})$) for the concept c_i evaluated on the mid-level feature $\mathbf{x}$, further normalized by a sigmoid function such that $0 < \varphi_i^V(\cdot) < 1$. The binary classifiers, that we name *visual classifiers* in the following, have been learned using images of the concept c_i as positive samples and images of a diversified class as negative samples. Each concept classification model $\varphi_i^V(\cdot)$ is obtained with L_2-regularized linear SVM, but other linear models could be used. Regarding the process of basic-level and subordinate concepts, even if it is similar, a particular difference is that, all basic-level concepts are selected in the final representation, while for subordinate concepts, we select only the most salients. This particular process for subordinate concepts avoid redundancy of information, due to the fact pointed in [20] that there is more concepts at a subordinate level than at the basic-level.

Concepts (c_i) at the highest categorical level (superordinate) are computed, for an input image, through a *semantic classifier*. It is an inference of concepts that have at least one hyponym relation with the superordinate concept (c_i). We thus, define the subsumption function that aims to output the set of concepts having hyponym relations with an input concept. We further, define the semantic classifiers that are used to compute superordinate concepts.

Definition 1. A subsumption function $\varsigma(\cdot)$ takes as input a concept c_i and a semantic hierarchy $\mathcal{H}$ with hyponymy relations and outputs a set $\mathcal{C}_i$ of concepts that are subsumed by the concept c_i, *i.e.*, the concepts that have an hyponymy relation with the concept c_i in a semantic hierarchy.

Definition 2. Considering $x \in \mathbb{R}^N$ a N-dimensional mid-level feature extracted from an image I. A semantic classifier is an operator that predicts the probability of presence of a concept c_i in the image through a semantic inference of purely visual output classifiers: $\varphi_i^S(\mathbf{x}, \mathcal{C}_i) = max(\varphi_{\mathcal{C}_1}^V(\mathbf{x}), \cdots, \varphi_{\mathcal{C}_M}^V(\mathbf{x}))$, where $\mathcal{C}_i$ is the set of concepts subsumed by c_i, $M = card(\mathcal{C}_i)$ and $\varphi^V(\cdot)$ is the output values given by visual classifiers.

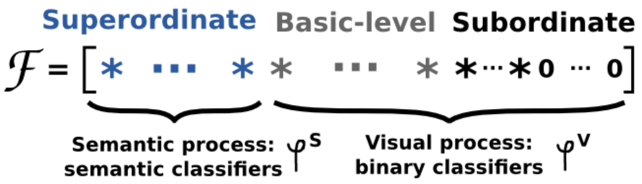
Superordinate
Basic-level
Subordinate
F = [* ··· * * ··· * *···*0 ··· 0]
Semantic process: semantic classifiers φS
Visual process: binary classifiers φV

Figure 3: Illustration of the asymmetric process in our D-CL feature. Superordinate concepts are processed semantically through semantic classifiers, while basic-level and subordinate concepts are visually processed through binary classifiers. Stars and zeros represents output values $\mathcal{F}_i(\cdot)$ of each concept of the D-CL feature $\mathcal{F}(\cdot)$. Note that, concepts are grouped by categorical levels, but any order could be obtained in a real scheme.

Finally, the proposed "Diverse Concept-Level" (D-CL) feature computes superordinate concepts through a semantic classifier and all other concepts, *i.e.* basic-levels and subordinates, using visual classifier. It also selects all basic-level and superordinate concepts and retains only the most salients subordinate concepts. Formally, let's $\mathcal{N}$ be the set of all concepts associated to a semantic hierarchy, $\mathcal{BL}$ the set of all basic-level concepts, $\mathcal{P}$ the set of superordinate concepts, $\mathcal{B}$ the set of subordinate concepts and $\mathcal{B}^K$ the set of the K most salient subordinate concepts for each input image. Note that, $\mathcal{N} = \mathcal{P} \cup \mathcal{BL} \cup \mathcal{B}$. Each dimension $\mathcal{F}_i(\mathbf{x})$ of the D-CL feature $\mathcal{F}(\mathbf{x})$ is a concept detector computed through:

$$\mathcal{F}_i(\mathbf{x}) = \begin{cases} \varphi_i^S(\mathbf{x}, \varsigma(i)), & \text{if } c_i \in \mathcal{P} \\ \varphi_i^V(\mathbf{x}), & \text{if } c_i \in \mathcal{BL} \cup \mathcal{B}^K \\ 0 & \text{if } c_i \in \mathcal{B} \setminus \mathcal{B}^K \end{cases} \tag{1}$$

where $\varsigma(\cdot)$ is the subsumption function, $\varphi_i^V(\cdot)$ the visual classifier, $\varphi_i^S(\cdot)$ the semantic classifier and K is a parameter corresponding to the number of subordinate concepts retained in the representation, that can be set by cross-validation. An illustration of the asymmetric process according to the type of concepts is presented in Figure 3.

3.2 Identifying Concept Groups in Practice

In this section, we detail how to identify the three groups of concepts (*i.e.*, basic-level, superordinate and subordinate), in practice, for a given supervised classification problem.

As depicted in Equation 1, the D-CL feature is computed by activating all the basic-level concepts, all the superordinate concepts, the K most salient subordinate concepts and by deactivating all others. Let $\mathcal{F}(\mathbf{x})$ be the D-CL feature of a mid-level feature $\mathbf{x}$ extracted for an image I contained in a targeted dataset. Let $\mathcal{D}^d$ the set of d categories of the targeted dataset.

While basic-level concepts are not available at a large scale, we propose to identify, in an offline phase, the set of basic-level concepts ($\mathcal{BL}$) selected in our D-CL feature by matching it with the set of targeted dataset categories $\mathcal{D}^d$. This latter, is based on the assumption that broader-datasets mostly contain categories at the basic-level. Specifically, all targeted dataset categories d_i are matched with concepts c_i to generate a set of basic-level concepts adapted to the dataset $\mathcal{BL}^d$. In fact, this matching has the advantage to make our D-CL feature adaptable to the application context. Regarding the sets of superordinate $\mathcal{P}$ and most salients subordinate $\mathcal{B}^K$ concepts, they are therefore automatically selected through the subsumption function $\varsigma(\cdot)$ that takes as input concepts from $\mathcal{BL}^d$ and a semantic hierarchy $\mathcal{H}$ with "is-a" relations. Formally, the Equation 1 becomes:

$$\mathcal{F}_i(\mathbf{x}) = \begin{cases} \varphi_i^S(\mathbf{x}, \varsigma(i)), & \text{if } c_i \in \mathcal{P}^d \\ \varphi_i^V(\mathbf{x}), & \text{if } c_i \in \mathcal{BL}^d \cup \mathcal{B}^K \\ 0 & \text{if } c_i \in \mathcal{B} \setminus \mathcal{B}^K \end{cases} \tag{2}$$

where $\mathcal{BL}^d$ and $\mathcal{P}^d$ are, respectively, the set of basic-level and superordinate concepts adapted to the targeted dataset $\mathcal{D}^d$. Selecting a portion of the whole concepts, and setting others to zero is closely related to the sparsification processes that sets to zero the lowest output values and keeps activated only the other concepts. Recent works underlined that such a property of sparsity has the advantage to be effective and computational efficient [10, 12]. The key novelty of our work is the adaptability of the concept selection to the input images. Contrary to former work, the sparsity is adapted to each image, according to its actual content, and relative to the problem of interest.

Our D-CL feature is illustrated in Figure 4. It is able to capture from an image containing multiple objects, all the basic-level concepts (colored in dark green) adapted to the target dataset, all its superordinate concepts (colored in dark red) and the most salient subordinate ones (colored in dark blue). It results in a final representation capturing the most informative concepts for a target collection of images.

4. EXPERIMENT AND ANALYSIS

In this section, we employ the proposed "Diverse Concept-Level" feature (denoted as **D-CL**) on three multi-object classification datasets. We first describe this datasets (Section 4.1) and the implementation details of our model (Section 4.2). Then, we report multi-object classification results on the three datasets (Section 4.3), and we compare it with the best semantic features in the literature. Finally, we evaluate the contribution of the asymmetrical process of concepts in the proposed D-CL descriptor by, first, evaluating the proposed semantic classifier and compare it to traditional binary classifiers (Section 4.4), and then, assessing the contribution of each concept groups selection (Section 4.5).

4.1 Datasets

The effectiveness of the proposed diverse concept-level feature is tested in the context of multi-class object classification. It is evaluated according to a standard experimental protocol as reported in the recent literature on the three following datasets:

- **Nus-Wide Object [4]**, is a multi-object classification dataset. As a subset of NUS-WIDE, it consists of 31 object categories and $36,255$ images in total. It contains $21,709$ images for training and $15,546$ images for testing. Each image is labeled by one or several labels from the 31 categories.

- **Pascal VOC 07 [8]** is a multi-object classification

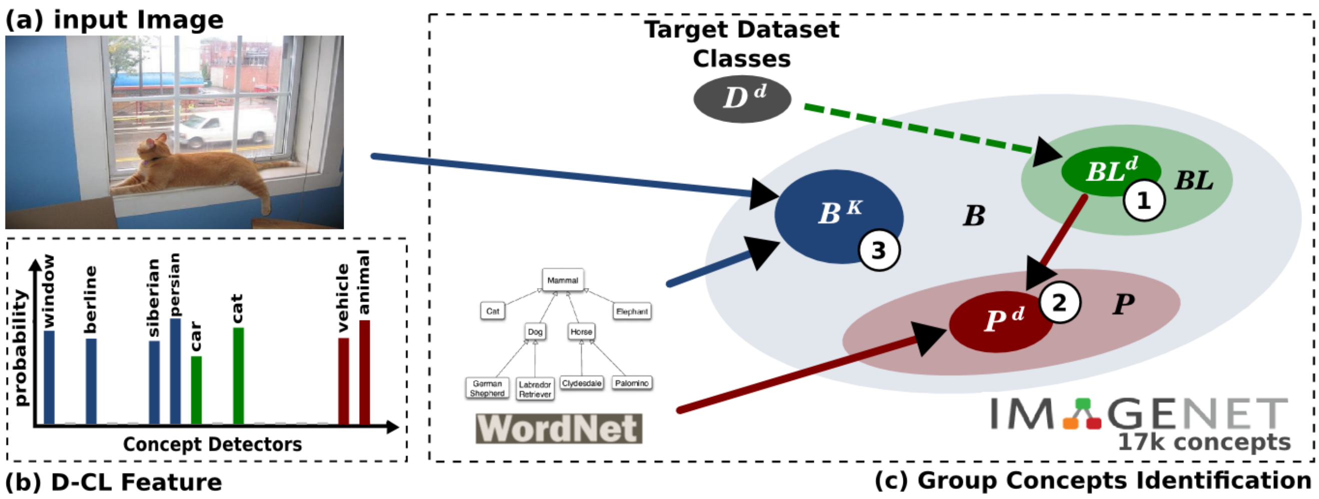

Figure 4: Illustration of the concept groups identification (c) in a practical case, for an input image (a) contained in a dataset collection. The proposed concept groups identification selects (1) in an offline phase (dashed arrow), the concepts of the target dataset categories ($\mathcal{D}^d$) as a portion ($\mathcal{BL}^d$) of all basic-level concepts ($\mathcal{BL}$), (2) the part ($\mathcal{P}^d$) of its superordinate concepts ($\mathcal{P}$) and in a final step (3) the most salient ($\mathcal{B}^K$) subordinate concepts ($\mathcal{B}$). For steps (2) and (3), a semantic hierarchy (WordNet) is used to compute the hyponymy relations. This latter, results in the final D-CL representation (b), to an activation of diverse concept levels (*i.e.*, superordinate, basic-level and subordinate) and a deactivation of all other concepts.

task. It is based on a dataset that contains 9,963 images, each image being labeled by one or several labels from 20 categories. We used the pre-defined split of 5,011 images for training and 4,952 for testing.

- **Pascal VOC 12 [7]** benchmark is similar to VOC 2007 but its number of images is larger: 22,531 images are split into 11,540 images for training and 10,991 images for testing.

4.2 Implementation details

D-CL learning: For all experiments, ImageNet [5] is used to learn our diverse concept-level representation. We especially use a subset of ImageNet with 17,462 concepts, containing more than 100 images each. Thus, we learn each individual concept detector using images representing the concept c_i as positive samples, and images of a diversified class as negative samples. Note that the concepts can be at any categorical-level of a semantic hierarchy, making our method applicable on top of any semantic feature.

Concept Groups Identification: As depicted in Section 3.2, the set of basic-level concepts ($\mathcal{BL}^d$ in Equation 2) is matched with the set of targeted dataset categories, for each dataset. Since all the concepts of ImageNet are organized in accordance to the WordNet [17] hierarchy, we use it as input to the subsumption function $\varphi(\cdot)$ to select the corresponding superordinate concepts ($\mathcal{P}^d$ in Equation 2). Specifically, only the first and the fourth level of the WordNet hierarchy are used. This, avoids redundancy of semantically close superordinate concepts. In fact, those levels contains the most popular superordinate concepts employed in cognitive experiments [15, 20, 25]. For the set of the K most salients subordinate concepts ($\mathcal{B}^K$), the parameter K of Equation 2, is cross-validated on each training dataset using the usual train/val split.

CNN feature: Semantic features (including the proposed D-CL), are built on top of any low-level or mid-level features (CNN). However, the quality of the D-CL feature will directly depend on the low/mid-level feature used. We thus, created semantic features on top of a competitive mid-level feature released in the literature, namely VGG-Net [24]. It is extracted from the last fully-connected layer (layer 16) of a Convolutional Neural Networks (CNN) learned on ILSVRC 2012 dataset [21] (containing 1.2 million images over 1,000 output categories), resulting in 4,096 dimensional vectors. Note that, for a fair comparison, the same mid-level feature is used to build Classemes+ [26] and Semfeat [10], presented in Section 4.3. 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 the available CNN models.

4.3 Multi-Object Classification Results

In this section, we test the D-CL feature for multi-object classification on the datasets presented in Section 4.1. The evaluation of our method lies in the context of semantic features. Thus we compare its performances to the following four baselines:

- **VGG-16 (fc8) [24]**, is extracted from a fully-connected layer (fc8, 18^{th} layer) of a CNN architecture (D) learned on ILSVRC 2012 dataset [21] that contains 1.2 million images of 1,000 classes. The resulting vector has 1,000 dimensions and can be seen as a semantic feature build on top of the fc7 (16^{th} layer), which the concept detectors are the final outputs of the CNN;
- **Semfeat [10]**, is built on top of a mid-level feature (Overfeat [23]) in their original work. To fairly com-

Method	Nus-Wide Object (20%)	Pascal VOC 2007 (45%)	Pascal VOC 2012 (30%)
ObjectBank [14]	n.a	45.2*	n.a
Classemes [26]	n.a	43.8*	n.a
Classemes+ [26]	70.3	82.4	81.7
Picodes [2]	n.a	43.7*	n.a
Meta-Class [1]	36.5	48.4 (53.2*)	49.3
VGG-16 (fc8) [24]	67.3	77.4	77.2
Semfeat [10]	74.7	82.8	81.7
D-CL (ours)	**76.0**	**85.1**	**83.0**

Table 1: Overall performance (mean Average Precision in %) of the following methods, ObjectBank, Classemes, Classemes+, Picodes, Meta-Class, VGG-16 (fc8), Semfeat and our approach (D-CL) on Nus-Wide Object, Pascal VOC 2007 and Pascal VOC 2012. We mention, for each dataset (in parenthesis), the rate of images labelled with multiple labels. Results marked with * are those reported in the original papers.

pare it to our method, we build it on top of the 16^{th} layer of VGG-16. This layer is used to learn the classifiers of the $17,462$ concepts of ImageNet that contains more than 100 images. According to their original work, a fixed sparsification over images is considered;

- **Classemes+** is, for a fair comparison with other methods, our own implementation of **Classemes [26]**. We build it on top of a the 16^{th} layer of VGG-16 with the same concepts as our method and Semfeat, that is to say $17,462$ concepts of ImageNet containing more than 100 images. Like in the original work, no sparsification is considered;
- **Meta-Class [1]**, is the output of $15,232$ concept detectors. 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 almost equal to other methods and the code is available, we use it as it is released [1].

To extend the comparison, we also report released scores by other semantic-based approaches in the literature (ObjectBank [14], Picodes [2] and Classemes [26]). Regarding the classification protocol, each class of the datasets is learned by a one-vs-all linear SVM classifier and we use mean Average Precision (mAP) to evaluate the performances. For each dataset, the cost parameter of the SVM classifier and the parameter K of Equation 2 are optimized through cross-validation on the training images, using the usual train/val split. Results are reported in Table 1. Our descriptor significantly outperforms all the other representations. On Pascal VOC 2007, D-CL has better performances than the four baselines Classemes+ (+2.7 points of mAP), Meta-Class (+35), VGG-16-fc8 (+7.7) and Semfeat (+2.3 points of mAP). Our method also outperforms the other semantic-based methods (ObjectBank, Picodes and Classemes) evaluated on Pascal VOC 2007. The same improvements are observed on Pascal VOC 2012 and Nus-Wide Object datasets. However, we note that, compared to all baselines, the improvements of the proposed D-CL feature, is much better on Pascal VOC 2007 than Pascal VOC 2012 and Nus-Wide Object. This result is aligned with the expectation since Pascal VOC 2007 contains a larger part (45%) of images labeled by multiple classes, compared to Pascal VOC 2012 and Nus-Wide Object, that contains only 30% and 20%, respectively. Regarding this, the performances of our method increases with the level of co-occurrence objects in the dataset and even achieves better performances than comparable state-of-the-art methods when the objects in the datasets have a lower level of co-occurrence.

4.4 Accuracy of Semantic Classifiers

In this section, we assess the effectiveness of the proposed semantic classifier ($\varphi^S(\cdot)$ of Equation 1), and compare it with purely visual classifiers, *i.e.*, binary classifiers ($\varphi^V(\cdot)$ of Equation 1) on generic concepts (*i.e.* concepts that have at least one hypernym relation with another concept).

This analytic study is an analogy to the experiment conducted by the cognitive works of Stephan Kosslyn [15]. More precisely, they wanted to provide a converging evidence that superordinate concepts are semantically processed by humans, rather than by a visual perception processing. Thus, to respect the analogy with [15], we evaluate the proposed semantic classifier and the visual classifiers on superordinate concepts only.

Regarding our experiment, the selection of superordinate concepts impose, in the Equation 2 of the proposed D-CL representation, to set to zero all the basic-level and subordinate concepts ($\varphi_i^V(\mathbf{x}) = 0, \forall c_i \in \mathcal{BL}^d \cup \mathcal{B}^K$). Thereby, the experiment has been conducted on the context of multi-class object classification through the Pascal VOC 07 dataset. All the images of the dataset have been re-labeled at superordinate level, *e.g.* all images labeled as *bird*, *dog*, *cow*, *horse* or *sheep* are now labeled as *animal*, all images labeled as *chair*, *sofa* or *table* are now labeled as *furniture*, etc (see first and second column of Table 2 for the re-labeling of other classes). Hence, we learn each superordinate class of the dataset by a one-vs-all SVM classifier. The cost parameter of the SVM classifier is optimized through cross-validation on the training dataset, using the usual train/val split. Performance results of both classifiers are reported in Table 2 using average precision (AP in %) for each class and mean Average Precision (mAP in %) over all classes in the last row. The second column of the table corresponds to the basic-level categories of the dataset and the first column

[1] http://vlg.cs.dartmouth.edu/projects/metaclass

Superordinate	**Basic-level**	**Visual**	**Semantic** (↑)
Animal	bird - cow - dog - horse - sheep	92.9	**97.7** (+4.8)
Electronic equipment	tv monitor	52.1	**72.6** (+20.5)
Furniture	chair - sofa - table	70.0	**74.9** (+4.9)
Person	person	77.2	**85.7** (+8.5)
Plant	potted plant	26.5	**40.5** (+14.0)
Vehicle	airplane - bike - boat - bus - car - mbike - train	93.4	**96.9** (+3.5)
Vessel	bottle	18.7	**31.4** (+12.7)
mAP		61.5	**71.4** (+9.9)

Table 2: Evaluating purely visual binary classifiers (denoted as "Visual") and our proposed semantic classifiers (denoted as "Semantic") for superordinate concepts (first column) of Pascal VOC 07 dataset classes (second column). The improvements of semantic classifiers over visual classifiers are shown in parentheses. Note that, the class *person* of Pascal VOC 2007 is already at the highest level in the WordNet hierarchy.

corresponds to the list of selected superordinate concepts in our D-CL feature. The average precision of each superordinate concept computed through binary classifiers (denoted as **Visual**) and the proposed semantic classifier (denoted as **Semantic**), are presented in the last two columns, respectively. Remarkably, the proposed semantic classifier clearly outperforms binary classifiers (purely visual) for all the superordinate concepts. From this study, we conclude that the superordinate concepts are better recognised by D-CL, due to its ability to compensate low within-category resemblance of generic concepts. The most surprising aspect of this experiment is that it shows, as concluded by the analogical cognitive experiment of Stephan Kosslyn [15], that semantic process is most adapted than purely visual process for superordinate concepts.

Concept Groups Selection	$\mathcal{P}$	$\mathcal{BL}$	$\mathcal{B}$	$\mathcal{B}^K$	**mAP**
Superordinate	✓				44.4%
Basic-level		✓			76.1%
Subordinate			✓		82.1%
K-Subordinate				✓	78.9%
Fusion 1	✓	✓	✓		82.7%
Fusion 2 (D-CL)	✓	✓		✓	85.1%

Table 3: Evaluation of the contribution of different concept groups selection (check-mark = selected group) in the proposed semantic feature on Pascal VOC 2007 dataset.

4.5 Concept Groups Selection Sensitivity

We evaluate now the contribution of the concepts from different groups (*i.e.* categorical levels) on a multi-object classification task (Pascal VOC 2007). To this end, we need to isolate each group of concepts in the D-CL representation by selecting them individually and setting other groups to zero. It results in four special cases of the D-CL feature (Equation 2), (i) selecting only superordinate concepts ($\forall c_i \in \mathcal{P}^d \cup \mathcal{B}, \varphi(c_i) = 0$) denoted as "Superordinate", (ii) selecting only basic-level concepts ($\forall c_i \in \mathcal{P}^d \cup \mathcal{B}, \varphi(c_i) = 0$) denoted as "Basic-level" (iii) selecting only subordinate concepts ($\forall c_i \in \mathcal{P}^d \cup \mathcal{BL}^d, \varphi(c_i) = 0$), denoted as "Subordinate" and (iv) selecting only the K most salient subordinate concepts ($\forall c_i \in \overline{\mathcal{B}^K}$), denoted as "$K$-Subordinate". We also evaluate the contribution when selecting all the concept groups in the representation ($\forall c_i \in \mathcal{P}^d \cup \mathcal{BL}^d \cup \mathcal{B}, \varphi(c_i) \neq 0$ in Eq. 2), *e.g.*, superordinate, basic-level and subordinate concepts, denoted as "Fusion 1". Finally, we report the results obtained by the proposed D-CL concept groups selection (see Section 3.2), corresponding to the selection of, all the superordinate and basic-level concepts and the K most salient subordinate concepts. It is also a fusion of other groups of concepts that we denote as D-CL. Results are reported in Table 3. For each concept group selection, a check-mark represents the concept groups that had been selected in the final representation. The last column gives the mAP obtained for the different concept selections. Note that, the K parameter of Equation 2 has been cross-validated for the "K-Subordinate" and "D-CL" concept group selections.

Obviously, selecting only superordinate concepts ($\mathcal{P}$) leads to very bad results, compared to basic-level concepts only ($\mathcal{BL}$), which are their-self lower than subordinate concepts only ($\mathcal{B}$). Selecting only the K most salient subordinate concepts ($\mathcal{B}^K$) obtains lower performances than selecting them all. Surprisingly, for the fusion, it is better with the selection of the K most salient concepts (the proposed D-CL) than with the selection of all subordinate concepts (Fusion 2). This experiment shows that the proposed D-CL selection gives a most effective semantic representation.

5. CONCLUSIONS

We propose the Diverse Concept-Level feature (D-CL), a semantic representation based on the exploitation of human knowledge, such as semantic hierarchies, to identify group of concepts, according to their categorical-level. This latter, aims to process the three groups of visual concepts differently from each other. Thus, our scheme outputs only informative concepts in the final representation. In addition, we show that the proposed semantic classifiers are most adapted to recognize superordinate concepts in images, than traditional visual classifiers. We also explored the selection of concepts from the three different categorical-levels, showing that the proposed scheme, consisting in the selection of concepts from all of them, is beneficial to obtain a precise semantic representation.

Experimental validation of the proposed approach has been conducted on three benchmarks (Pascal VOC 2007, Pascal VOC 2012 and Nus-Wide Object) of multi-class object classification. The proposed D-CL feature obtained significantly better performances than the best semantic features in the literature.

The results obtained for image classification are very encouraging and we will pursue the work reported here. We will investigate finer ways to identify basic-level concepts. In particular, large released lists of basic-level concepts [16, 19, 22] will replace the dataset categories that are currently used. This work direction, will aim to handle unsupervised image retrieval problem, where categories of images in the collection are not supposed known.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

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

[2] A. Bergamo, L. Torresani, and A. W. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*, 2011.

[3] 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.

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of ACM Conference on Image and Video Retrieval*, CIVR, 2009.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[6] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Computer Vision and Pattern Recognition*, 2012.

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

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

[9] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Computer Vision and Pattern Recognition*, CVPR, 2009.

[10] 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.

[11] Y. Huang, Z. Wu, L. Wang, and T. Tan. Feature coding in image classification: A comprehensive study. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(3):493–506, 2014.

[12] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *International Conference on Computer Vision*, ICCV, 2015.

[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 *Advances in Neural Information Processing Systems*, NIPS, 2010.

[15] P. Jolicoeur, M. A. Gluck, and S. M. Kosslyn. Pictures and names: Making the connection. *Cognitive Psychology*, 16(2):243–275, 1984.

[16] A. Mathews, L. Xie, and X. He. Choosing basic-level concept names using visual and language context. In *Winter Conference on Applications of Computer Vision*, WACV, 2015.

[17] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[18] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. Berg. From large scale image categorization to entry-level categories. In *International Conference on Computer Vision*, ICCV, 2013.

[19] V. Ordonez, W. Liu, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. Predicting entry-level categories. *International Journal of Computer Vision (IJCV)*, pages 1–15, 2015.

[20] E. Rosch. Principles of categorization. *Cognition and Categorization*, page 2748, 1978.

[21] 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.

[22] 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 (IJCV)*, 115(3):211–252, 2015.

[23] 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.

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

[25] J. W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482, 1991.

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

[27] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, series B*, 68:49–67, 2006.