MuCaLe-Net: Multi categorical-Level networks to generate more discriminating features

Y. Tamaazousti, Hervé Le Borgne, Céline Hudelot

MuCaLe-Net: Multi Categorical-Level Networks
to Generate More Discriminating Features

Youssef Tamaazousti[1,2]   Hervé Le Borgne[1]   Céline Hudelot[2]

[1]CEA, LIST, F-91191 Gif-sur-Yvette, France.

[2]CentraleSupélec (University of Paris-Saclay), MICS, 92295 Châtenay-Malabry, France.

Abstract

In a transfer-learning scheme, the intermediate layers of a pre-trained CNN are employed as universal image representation to tackle many visual classification problems. The current trend to generate such representation is to learn a CNN on a large set of images labeled among the most specific categories. Such processes ignore potential relations between categories, as well as the categorical-levels used by humans to classify. In this paper, we propose Multi Categorical-Level Networks (MuCaLe-Net) that include human-categorization knowledge into the CNN learning process. A MuCaLe-Net separates generic categories from each other while it independently distinguishes specific ones. It thereby generates different features in the intermediate layers that are complementary when combined together. Advantageously, our method does not require additive data nor annotation to train the network. The extensive experiments over four publicly available benchmarks of image classification exhibit state-of-the-art performances.

1. Introduction

Convolutional neural networks (CNNs) established the state-of-the-art for several visual recognition problems. One of the most important reasons behind their success is their ability to generate rich discriminating features (convolutional and unit-filters). The common scenario to obtain such performing features is to solve a *discriminative problem*, that consists to separate categories from one another. By solving this discriminative problem, features are *generated* at the intermediate layers of the network. In a transfer-learning scenario [2, 23, 38], these latter can then serve as image representation for a new *target problem* with few training-data. Since these CNN-based descriptors have obtained good performance on many target-problems, they were assigned the property of *universality*. An important question, that the current article investigates, is thus to determine how to obtain *more universal* image representations resulting from the *generative aspect* of a CNN.
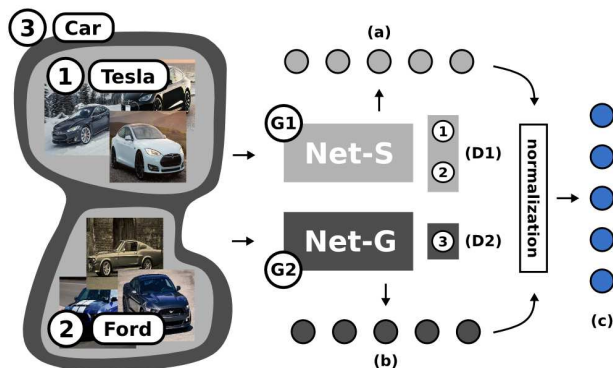
The problem of increasing the universality of a CNN-

Figure 1. A standard CNN (Net-S) solves a discriminative problem (D1) that consists to separate specific categories ((1) and (2)) from each other. This last, automatically generates powerful specific features in the intermediate layers (G1), from which the fully-connected activations (a) are generally used as image-representation in a transfer-learning scheme. Here, we advocate that varying the discriminative problem aims to vary the generated features, even through the same training-images. More precisely, our method re-labels the training-images at a generic categorical-level (simply by re-labeling the categories), then use these generic categories (3) as the discriminative problem (D2) to be solved by another CNN (Net-G) for generating different features (G2). After normalization, we combine the specific (a) and generic (b) fully-connected activations, to get the final MuCaLe representation (c). In contrast to standard CNN-descriptors, MuCaLe is more diversified in terms of relevant features (convolutional and unit-filters) leading to better performances. Best view in color.

descriptor can be handled by increasing the quantity of relevant filters generated by a CNN. However, it is unclear how to directly supervise their complex generative aspect. Consequently, many works indirectly tackle the issue by focusing on the discriminative problem. In particular, two approaches emerge when reviewing the literature. The first one is known as the "ensemble-model" [14, 15, 31, 36, 40], and the second as the "data-enlargement" [2, 17, 21, 41]. The former consists to train many different networks and merge their internal features as image representation to solve the target-problem. The difference between each network can be the random initialization [40], the image-

scale [14, 31] or more recently the subset of categories to recognize [1, 15, 22, 27, 36]. However, this approach needs many cumbersome models to get sufficient diversity in the ensemble, which is highly limited on a test phase when time is critical. The second approach (data-enlargement) consists to enlarge the training-database through more images-per-class and more categories. By this enlargement, the CNN is forced to solve a different discriminative problem that desirably increases the universality of generated features. However, the key limitation of this approach is its cost, since it requires thousands additive images and their corresponding annotations. Furthermore, commonly to all previously mentioned methods, only specific categories are considered in the discriminative problem whereas it has been shown by psychologists such as Eleanor Rosch [29] and Stephan Kosslyn [16] that humans use many categorical-levels[1] to classify. For this reason, recent works in computer vision highlighted the importance to pay attention to the exact way we name objects [6, 20, 25, 26, 34].

In this paper, we propose a new method to relevantly increase the diversity of the generated features based on the nesting of Human-categorization [16, 29] (*e.g.*, categorical-level labels of objects) into the CNN learning process. Our proposal belongs to the ensemble-model approach with a new original definition of the difference between the networks. This difference is obtained from a variation of the categorical-level labels of the training-images. More precisely, our main contribution consists to structure the labels of the training-images in multi categorical-levels and independently train one network per categorical-level, resulting to a new learning-strategy that we name *MuCaLe-Net*. For new images, we extract the intermediate layers from the pretrained MuCaLe-Net and merge them to get the final *MuCaLe* representation. Figure 1 illustrates the whole method. To the best of our knowledge, this is the first attempt to use multi categorical-level networks to obtain a more universal image representation in a transfer-learning scenario.

A major advantage of the proposed method in comparison to previous work is that, is does not require any additive image dataset nor very costly manual annotation, while it achieves state-of-the-art performances on four publicly available benchmarks in image classification.

Above the proposal itself and the demonstration of it performance in practice, the other major contribution of this paper lies in Section 4 in which we analyze and clarify the reasons why our approach works. Compared to representations obtained from standard CNNs trained with specific labels, an advantage of MuCaLe appears when the filters fail at the subordinate-level (*e.g.* in Fig 1, the filters for *Tesla*

and *Ford* are both weakly activated), which is often the case since the categories are finer thus harder to identify. With our proposal, the descriptor at least contains features that capture common properties among basic-level categories (*e.g.* filters of *Car* are highly activated), making it more robust for classification problems.

2. Related Work

Diversification-strategies. The data-enlargement approach that consists to add images labeled among new categories (specific [2, 41], generic [18, 21], noisy tags [17]) to the initial discriminative problem is a powerful diversification-strategy. A strong limitation of these methods lies in their cost resulting from the need for many additive data and corresponding annotations. The works of [18, 21] are the closest to ours, since they consider both generic and specific categories. The key difference is the way they combine both types of categories, as well as the genericity definition. In fact, they solve only one discriminative problem by jointly training the CNN on the generic and specific data, resulting into a mix of generic and specific features in the intermediate layers. Moreover, the joint learning makes the generic categories (*e.g.* dog) and specific ones (*e.g.* rottweiler) mutually exclusive, which clearly violates real world semantics. In contrast, we independently solve two discriminative problems, resulting into a desirably clear separation of the different features. Moreover, they define the generic categories as the internal nodes of a hierarchy while we consider they belong to the basic categorical-level.

The ensemble-model approach [1, 15, 22, 27, 36, 40] is another way of diversifying the features. Desirably, it does not need additive training-data, but it requires many networks to diversify the ensemble. From this approach, the work of [27] is the closest to ours. They use "abstract" generic categories (do not exist in the real world) that capture common properties among many object classes. They are built using hierarchical clustering on low-level features of images among the initial set of categories. The restrictive assumption of this method is its dependency to the visual low-level features, since it leads to irrelevant categories when low-level features fail to capture the dissimilarity between different categories. Moreover, they train a network on the initial set of categories and fine-tune it on all other generic groups, thus all new filters are highly biased by those of the initial model which is undesirable for increasing the set of features. In fact, they need 18 models to get significant diversity which is very costly. On the contrary, the method we propose relies on the human categorization expertise to reflect complex relations between categories. By independently training each discriminative problem, it only needs two models to significantly diversify the ensemble.

Cognitive studies in computer vision. A last line of work deals with the inspiration from cognitive studies in computer vision [6, 20, 25, 26, 34]. Generally, the main goal of

---

[1]Humans use *superordinate*, *basic-level* and *subordinate* categories. *Basic-level* categories correspond to the most common words used by Humans to categorize objects. *Superordinate*/*subordinate* categories correspond to categories that are more generic/specific than *basic-level* ones.

these works is to output the corresponding basic-level concepts of an image from a set of predicted concepts. In particular, the work of Deng *et al.* [6] is closest to ours since they optimize the trade-off between accuracy and specificity. In other words, if the concept detectors fail to recognize categories at a specific level, they try to return a more general concept. As in our work, their system reflects the psychological hint stating that, even if humans tend to categorize objects at subordinate level, they are still aware of the other categorical-levels. However, the key difference with our work is the method used to integrate this psychological hint as well as its purpose. In fact, our goal is to diversify the generated features in CNNs, while they aim at annotating the images and thus identifying one concept. Moreover, we opt for an integration of the psychological hint at three levels, the training-data, the learning process and the image representation design, while they do it only during the test phase, after the prediction of the different categories.

3. Proposed Approach

In this section, we aim at learning more universal image representations by nesting human-categorization knowledge into deep CNNs. Our approach consists of three modules: the re-labeling of categories (Sec. 3.1), the learning of multi categorical-level networks "MuCaLe-Net" (Sec. 3.2) and the extraction of the multi categorical-level "MuCaLe" representation (Sec. 3.3). An overview of the whole approach is illustrated in Figure 3.

3.1. Categorical-Level Re-labeling

Let us consider a semantic hierarchy with hyponymy relations, that is to say, a (large) set of categories organized according to "is-a" relations. An example of such a hierarchy is WordNet, on which are mapped the categories of ImageNet [5]. Formally, this is a directed acyclic graph $\mathcal{H} = (\mathcal{V}, E)$ consisting of a set $\mathcal{V}$ of nodes and directed edges $E \subseteq \mathcal{V} \times \mathcal{V}$. Each node $v \in \mathcal{V}$ is a label and an edge $(v_i, v_j) \in E$ is a hierarchy-edge indicating that label v_i subsumes label v_j. Let us also consider a dataset $\mathcal{D}_\mathcal{C}^N$ containing N images labeled among S specific categories belonging to $\mathcal{C} = \{c_1, c_2 \ldots, c_S\}$, such that $\mathcal{C} \subset \mathcal{V}$.

We now consider a partition of $\mathcal{C}$ into G subsets *i.e* $\mathcal{C} = \bigcup_{i=1}^{G} \mathcal{C}_i$ (an example to obtain such a partition is detailed in supp. material) and define a re-labeling function as:

$$R: \quad 2^\mathcal{C} \rightarrow \mathcal{V} \qquad (1)$$
$$\mathcal{C}_i \mapsto LCA_\mathcal{H}(\mathcal{C}_i),$$

where $LCA_\mathcal{H}(\mathcal{C}_i)$ is the lowest common ancestor of the categories of $\mathcal{C}_i$ according to $\mathcal{H}$. Let us note that the partition of $\mathcal{C}$ is the *only* additive work needed. The "cost" of this step is thus much smaller than that usually needed, *i.e*, collecting new images and annotating each of them. Using this function R, we obtain G categories, with $G \ll S$.
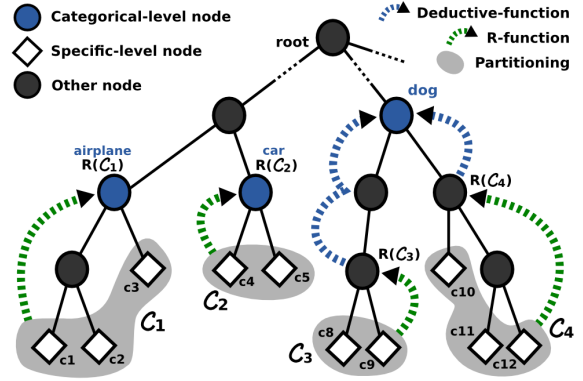

Figure 2. Illustration of our categorical-level re-labeling. Given a set of categories (leaf nodes of the hierarchy in white diamonds) and a partitioning of it into G subsets $\mathcal{C}_1, \cdots, \mathcal{C}_G$ (gray blobs), our method automatically re-labels them (and thus, their images) into generic categories (blue circles). It consists to first get the least common ancestor of each subset $\mathcal{C}_i$, through the R-function (green dashed arrows) then, if it does not reach a category of the desired categorical-level $\mathcal{L}$ (due to the hierarchy-imbalance), it applies a deductive function (blue dashed arrows) that goes to its ancestors until it reaches it. Best view in color.

We now consider a categorical-level defined according to human cognition. Let us note $\mathcal{L}$ a set of categories that belong to a categorical-level [16, 29, 35] (*e.g*, *basic* or *superordinate*-levels). It is important to realize that the categories of $\mathcal{L}$ do *not* correspond to a given level of the hierarchy $\mathcal{H}$. Our purpose here, is to match the G categories previously re-labeled by Eq (1) to $\mathcal{L}$. In fact, as illustrated in Fig. 2, some of the G categories already match the desired categorical-level $\mathcal{L}$ but others may not (due to the imbalance of semantic hierarchies). For this, we consider a *deductive function* $\delta_\mathcal{H}(\cdot)$ that associates to a category v_i of $\mathcal{V}$ its direct ancestor, that is to say, the category directly above v_i according to $\mathcal{H}$. We note $\delta_\mathcal{H}^n(\cdot)$ the corresponding iterated function (*i.e* $\delta_\mathcal{H}(\cdot)$ composed with itself n times) and we assume that the image of the root node of $\mathcal{H}$ is itself. Hence, given a category c_i, the set of all its ancestors in $\mathcal{V}$ is $\mathcal{A}_\mathcal{V}(c_i) = \{\delta_\mathcal{H}^j(c_i)\}_{j=1}^{\infty}$. We can now define our re-labeling function relative to a given categorical-level $\mathcal{L}$ by:

$$R_\mathcal{L}: \quad 2^\mathcal{C} \rightarrow \mathcal{L} \qquad (2)$$
$$\mathcal{C}_i \mapsto \mathcal{L} \cap \mathcal{A}_\mathcal{V}(LCA_\mathcal{H}(\mathcal{C}_i)).$$

Simply said, while Eq. (1) identifies the least common ancestor of the subsets of $\mathcal{C}$, Eq. (2) matches their ancestors belonging to a categorical-level $\mathcal{L}$ (illustration in Fig. 2).

Thus, given the initial training-dataset $\mathcal{D}_\mathcal{C}^N$ (labeled according to $\mathcal{C}$) and the hierarchy $\mathcal{H}$, the only additive work is to determine the set of categories in $\mathcal{L}$. Hence, assuming that none of them is an ancestor of an other in $\mathcal{V}$ (*i.e* $\forall(v_i, v_j) \in \mathcal{L} \times \mathcal{L}, v_i \notin \mathcal{A}_\mathcal{V}(v_j)$) it is straightforward to automatically re-label $\mathcal{D}_\mathcal{C}^N$ according to $\mathcal{L}$, once it is chosen.

3.2. Multi Categorical-Level Network Learning

Given the initial training-dataset $\mathcal{D}_{\mathcal{C}}^N$ (with specific labels) and the ω re-labeled generic ones $\{\mathcal{D}_{\mathcal{L}_1}^N, \cdots, \mathcal{D}_{\mathcal{L}_\omega}^N\}$, we now learn the Multi Categorical-Level Network (**MuCaLe-Net**). Generally, it consists to learn one CNN per database (initial database $\mathcal{D}_{\mathcal{C}}^N$ and re-labeled ones $\mathcal{D}_{\mathcal{L}_i}^N$) in accordance to the classical methodology [18]. An illustration is given in Figure 3. Note that our method does not depend on a particular network architecture, thus can benefit from the advances in this domain.

Formally, let us consider a training-database $\mathcal{D}_{\mathcal{B}}^N$ that contains N images $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ labeled among the categories of the label-set $\mathcal{B} = \{b_1, \cdots, b_{card(\mathcal{B})}\} \subset \mathcal{V}$. Let $\mathcal{B} = \mathcal{C}$ for the initial database $\mathcal{D}_{\mathcal{C}}^N$, and $\mathcal{B} = \mathcal{L}_i$ for those re-labeled according to categorical-levels $\mathcal{D}_{\mathcal{L}_i}^N$. The multiple datasets could be differentiated by other principles than *categorical-levels*. Using the "softmax" to specify how to penalize the deviation between the predicted and true labels, the posterior probability of an image $\mathbf{x}_i$ and category b_j for the label-set $\mathcal{B}$ is:

$$p_{ij}^{\mathcal{B}} = \frac{\exp(f_j^{\mathcal{B}}(\mathbf{x}_i))}{\sum_{k=1}^{Card(\mathcal{B})} \exp(f_k^{\mathcal{B}}(\mathbf{x}_i))}, \qquad (3)$$

where $f_j^{\mathcal{B}}(\mathbf{x}_i)$ is the jth dimension of the output of the last fully-connected layer of the network and the dimensionality of $f^{\mathcal{B}}(\cdot)$ is equal to the number of categories in label-set $\mathcal{B}$. Thus, assuming that the ground-truth probability for image $\mathbf{x}_i$ and class b_j at categorical-level $\mathcal{B}$ is defined as $\overline{p_{ij}^{\mathcal{B}}}$, the cost function (from maximum log-likelihood) to be minimized by asynchronized stochastic gradient descent is:

$$J_N^{\mathcal{B}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{Card(\mathcal{B})} \overline{p_{ij}^{\mathcal{B}}} \log(p_{ij}^{\mathcal{B}}). \qquad (4)$$

Note that, we have as much cost functions to minimize as the number of label-sets $\mathcal{B}$, and each cost function $J_N^{\mathcal{B}}(\theta)$ is minimized independently from all others, on the same N images of the initial database $\mathcal{D}_{\mathcal{C}}^N$. At convergence, we obtain a set $\Omega = \{W_{\mathcal{C}}, W_{\mathcal{L}_1}, \cdots, W_{\mathcal{L}_\omega}\}$ of $\omega + 1$ models, with ω corresponding to the number of re-labeled label-sets used. Note that, by considering the categorical-levels to differentiate the databases, we can re-label the initial label-set to the *basic* and the *superordinate* categorical-levels. Nevertheless, the proposed formalism for multi-level network is general, and other principles may drive the definition of each database-level.

In practice, for the initial training-database $\mathcal{D}_{\mathcal{C}}^N$, we use the popular ILSVRC dataset [5], which is a subset of ImageNet. Its set of categories corresponds to the most specific hierarchical-level of ImageNet (*i.e*, the most specific words used by humans with domain-expertise to categorize objects) thus, it can be considered as the *subordinate*-level.
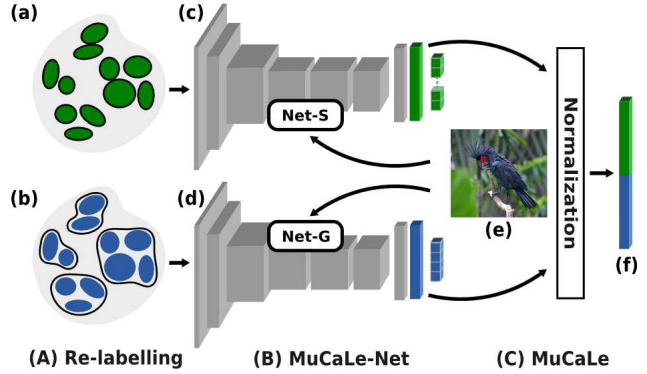

Figure 3. Illustration of the MuCaLe-Net learning procedure (B) and the extraction of the MuCaLe descriptor (C) given the different categorical-level databases obtained after the re-labeling procedure (A). MuCaLe-Net consists to individually train one network per categorical-level database (that contains the same images but differs by their label-sets). Here, we have two databases (specific (a) and generic (b)). The network trained on the specific database is denoted by Net-S (c) and the one trained on the generic one, by Net-G (d). For a new image (e), we individually extract one of the layers from each component of the pre-trained MuCaLe-Net, independently-normalize and concatenate them to obtain the final MuCaLe representation (f). Best view in color.

Hence, using categorical-levels as a principle to differentiate the multiple databases, we have $\mathcal{B} \in \{\mathcal{C}, \mathcal{L}_1, \mathcal{L}_2\}$ with $\mathcal{C}$, $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively corresponding to *subordinate*, *basic* and *superordinate* categorical-levels. After training our MuCaLe-Net with respect to Eq. (4), we obtain a set of three network-models $\Omega = \{W_{\mathcal{C}}, W_{\mathcal{L}_1}, W_{\mathcal{L}_2}\}$.

3.3. MuCaLe Representation

At testing time, let us consider a query image $\mathbf{x}_i$, and a set $\Omega = \{W_{\mathcal{C}}, W_{\mathcal{L}_1}, \cdots, W_{\mathcal{L}_\omega}\}$ of network models learned on training-databases labeled according to the initial label-set $\mathcal{C}$ and the ω re-labeled ones $\{\mathcal{L}_1, \cdots, \mathcal{L}_\omega\}$. Let us denote the feature extracted from a CNN model $W_{\mathcal{B}}$ by $\phi_K^{\mathcal{B}}(\cdot)$ where the Kth first layers filter the images (*e.g*, $\{conv1, conv2\}$ when $K = 2$ with AlexNet [18] network). The output of $\phi_K^{\mathcal{B}}(\cdot)$ is thus a scoring function of the data point $\mathbf{x}_i$ that produces a vector of activations. The Multi Categorical-Level (**MuCaLe**) representation for the query image $\mathbf{x}_i$ is thus computed as:

$$\Phi(\mathbf{x}_i, \Omega, K) = \bigoplus_{\mathcal{B} \in \{\mathcal{C}, \mathcal{L}_1, \cdots, \mathcal{L}_\omega\}} \left(\frac{\phi_K^{\mathcal{B}}(\mathbf{x}_i)}{\|\phi_K^{\mathcal{B}}(\mathbf{x}_i)\|_\infty} \right), \qquad (5)$$

where $\bigoplus_{\mathcal{B}}(\cdot)$ is the *concatenation* operator among the $\omega + 1$ input vectors, and $\|v\|_\infty$ returns the maximal value of v. In practice, when Ω contains the models obtained from the three categorical-level databases, Eq. (5) concatenates three features to get the MuCaLe representation. Fig. 3 illustrates the learning procedure (presented in previous section) and the extraction of the representation for a new image.
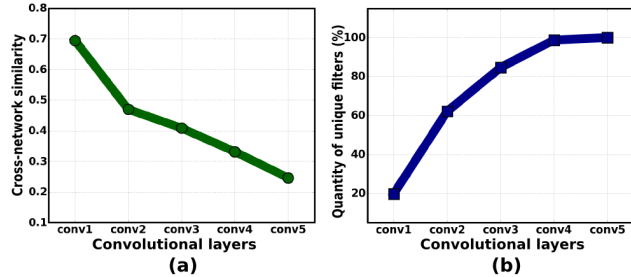
Figure 4. Graph (a) reports the average of maximum correlation between convolutional-filters of Net-G and Net-S (*i.e*, similarity between the two networks) according the layers. In (b), we plot the average quantity of unique filters (that do not match with any filter of the other network) in the two networks according the layers.

4. Diversification Ability of MuCaLe-Net

In this Section, we investigate whether each component – network trained on images labeled at a particular set of categorical-level labels – of a MuCaLe-Net introduces some relevant diversity in the whole features' space. To do so, we analyze (in Sec. 4.1) the intermediate layers of the different pre-trained categorical-level networks in order to show that different features have been generated between them. However, since diversity does not mean relevance, we also analyze (in Sec. 4.2) how pertinent is this difference.

4.1. Do different categorical-level networks learn different features?

Two categorical-level networks are trained using the AlexNet network on the same training-images (ILSVRC$^{0.5}$ detailed in Sec. 5.1) for two categorical-level labels (*basic* and *subordinate*), that we respectively denote by Net-G and Net-S. Following Li *et al.* [19], we statistically compare the internal convolutional-filters of the two networks. Thus, we first aggregate certain statistics of the activations within the networks. Given a pre-trained network Net-n, we denote by the scalar random variable $\mathcal{X}_{l,i}^{(n)}$ the activation values produced over a large set of samples[2] by convolutional-filter i on layer $l \in \{conv1, \cdots, conv5\}$. From this set of samples, we collect for all filters, the average $\mu_{l,i}^{(n)}$, the standard deviation $\sigma_{l,i}^{(n)}$ and the "cross-network correlation":

$$c_{l,i,j}^{(n,m)} = \mathbb{E}[(\mathcal{X}_{l,i}^{(n)} - \mu_{l,i}^{(n)})(\mathcal{X}_{l,j}^{(m)} - \mu_{l,j}^{(m)})]/\sigma_{l,i}^{(n)}\sigma_{l,j}^{(m)}, \quad (6)$$

corresponding to the correlation of the random variable of a filter of a network Net-n with the one of another network Net-m at the same layer l. We thus obtain five asymmetric matrices per network, of size 96×96 for $conv1$, 256×256 for $conv2$ and $conv5$ and 384×384 for $conv3$ and $conv4$. From these matrices, we look for the filters of Net-S that

---

[2]All the spatial positions (55×55 for $conv1$, 27×27 for $conv2$ and 13×13 for all others) of $2,000$ randomly selected images from the ILSVRC$^{0.5}$ validation set.
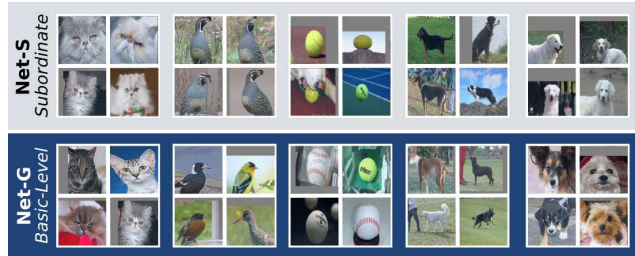

Figure 5. Illustration of the top-4 patches that highly activate some unique convolutional-filters of the $conv5$ layer. The top part reports unique filters learned from **Net-S** and the bottom part, unique filters from **Net-G**. More visualizations in supplementary material. Best view in color.

match the most with those of Net-G (and inversely) to show the similarity between the networks. To show the amount of unique filters generated in each network, we look for the filters that do not have any matching filter in the other network. Regarding the cross-network similarity, we compute the average of maximum cross-network correlation (*e.g.*, the average of the maximum of each row and each column of the cross-network correlation matrices), for each layer and display the results on the graph (a) of Figure 4. As in [19], for the amount of unique filters, we compute the relative percentage of filters that do not match with any filter in the other network – *e.g.*, the maximum cross-correlation of that filter with all those of the other network is above a low threshold of 0.5 – for all the layers and report the results on the graph (b) of Figure 4.

First of all, the high similarity (0.71) observed in $conv1$ confirms previous works [19, 38, 40] showing that, whatever the training-database, the first layers always generate very similar filters (blob and Gabor-like filters). Second, at the other extremity ($conv5$), the cross-network similarity is much lower (0.26) and the quantity of unique filters reached 100%, meaning that a very different feature space has been generated by the two networks. Hence, these two criteria show that two categorical-level networks begin by generating many common filters but end by learning many unique ones, even if the same training-images have been used. Equally important, graph (4b) shows that we have a difference of 20% between the two networks at $conv1$, clearly meaning that the representational divergence of MuCaLe-Net begins at the first layer and thus confirms the importance of training the networks independently without sharing any layer between them.

To enhance this study, we visualize some unique filters learned from each network at the last layer ($conv5$). To do so, we follow the literature [11, 39, 40]. In particular, for each convolutional-filter, we display in a two by two block the top-four image-patches (extracted from the large set of ILSVRC validation images) that highly activate them. Figure 5 shows those unique filters. More visualizations of unique filters are reported in supp. material, especially

for the other layers. We clearly observe that Net-S has learned specific filters such as breed of objects (*e.g.*, they highly activate only *tennis-ball*, *hairy-cat*, or even *white-dog* patches) while in the Net-G, generic filters that are invariant to the breed of objects have been learned (*e.g.*, they highly activate *cat*, *bird*, or *ball* patches). This visualization confirms that different filters are generated by different categorical-level networks.

This analysis shows that solving different discriminative problems (through different categorical-level labels) with a CNN forces the generation of different convolutional-filters. This latter, is quite surprising since the same training-images are used in the two discriminative problems.

4.2. Is the difference between each set of categorical-level features relevant?

In the previous section, we have shown that each component of a MuCaLe-Net generates different features. However, difference does not mean relevance, thus, it is crucial to evaluate whether this diversity in the feature's space becomes relevant in the final representation. For evaluating the relevance of a descriptor, we follow Peng *et al.* [28] and Herranz *et al.* [14] that estimate the relevance of a set of features by its discriminability on a set of categories. Formally, we define the *discriminability* of a N-dimensional representation $\Phi(\mathbf{x}) = \{\varphi_1(\mathbf{x}), \cdots, \varphi_N(\mathbf{x})\}$ with respect to a class b_j belonging to a set of M classes $\mathcal{B} = \{b_1, \cdots, b_M\}$ as

$$D(\Phi(\mathbf{x}), b_j) = \frac{1}{N} \sum_{i=1}^{N} I(\varphi_i(\mathbf{x}), b_j), \qquad (7)$$

where $I(\varphi_i(\mathbf{x}), b_j)$ is the mutual information of filter $\varphi_i(\mathbf{x})$ and class b_j. The filter $\varphi_i(\mathbf{x})$ is a continuous variable, thus, as in [28], we use a density estimation method (*e.g.*, Parzen windows) to approximate $I(\varphi_i(\mathbf{x}), b_j)$.

In Figure 6, we plot the discriminability on the 31 categories of the Nus-Wide Object dataset (details in Sec. 5.1) of the $fc7$ representations generated by three networks: (i) Net-G, (ii) Net-S, and (iii) MuCaLe-Net. Note that, the MuCaLe descriptor obtained by (iii) corresponds to the combination of those obtained by the first two. From this graph, we observe that the representation generated by Net-S is roughly more discriminating and thus different than the one learned by Net-G, confirming the results of the previous section. More interestingly, the graph shows that their combination (MuCaLe), is more discriminating (for all the categories) than each of them independently, meaning that the difference between the features is highly relevant.

To resume, this analysis shows that CNNs are able to generate different features with the same training-images by varying their categorical-level labels and that this difference is highly relevant. This means that the proposed MuCaLe-Net strategy is going on the direction of our main goal, that consists to diversify the set of features generated by CNNs in order to result in a more universal image representation.
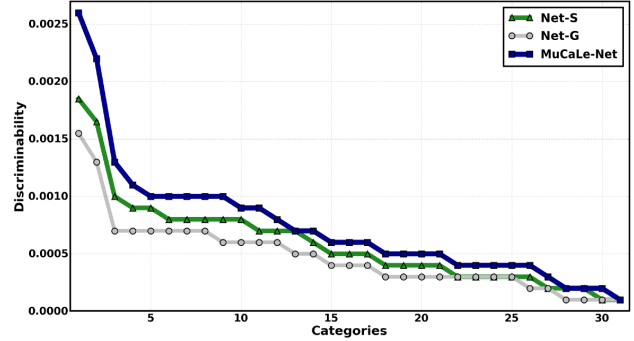

Figure 6. Comparison of the discriminability (ordered by decreasing values) of representations generated by Net-G, Net-S and MuCaLe-Net, on the categories of the NWO dataset.

5. Experimental Results

After describing the experimental settings and the datasets (Section 5.1), we compare (Section 5.2) the diversification performances of the proposed strategy with those of the literature. Then, we compare (Section 5.3) the performances of the MuCaLe representation with state-of-the-art CNN-based descriptors on an image classification task.

5.1. Experimental Settings

Network training. We train the networks using two popular architectures, AlexNet [18] and VGG-16 [32]. Since VGG is much deeper than AlexNet, it is technically hard to get convergence with full back propagation as for AlexNet. Thus, following the original paper [32], we initialize the first layers with those of a pre-trained network and re-train the whole network on the training database (more details in supplementary material). In fact, for the training-databases, we use two subsets of ImageNet, (i) ILSVRC$^{0.5}$: a large and clean set released in [30], containing half a million images labeled among 483 fine-grained categories and (ii) ILSVRC containing 1.2 million images labeled among $1,000$ fine-grained categories. Since the process is faster, (i) is used when many network-trainings are necessary.

Target-datasets. We used five popular object-recognition datasets, namely Pascal VOC 2007 [8] (**VOC07**), Pascal VOC 2012 [9] (**VOC12**), Caltech-101 [10] (**CA101**), Caltech-256 [12] (**CA256**) and Nus-Wide Objects [4] (**NWO**). We follow standard protocols for all the benchmarks, that we report in details in the supp. material.

Transfer-learning. We apply the common scenario that consists to consider a pre-trained network (on a large image database) as an image representation extractor and, each image is represented by one layer of the pre-trained network. Each class of the target-dataset is then learned by a *one-vs-all* SVM classifier. Fine-tuning could also be used to learn the target-classes but it will result in the adaptation of the representation to the target-dataset (*i.e*, domain adaptation), which is out of the scope of the paper. The performances are evaluated either with mean Average Precision (mAP) or

Strategy	VOC07	NWO
Standard	70.3	51.2
Random Difference (RD)	70.5±0.3	51.3±0.2
Random Grouping (RG)	70.6±0.1	51.0±0.3
Joint Training (JT)	71.5	53.4
MuCaLe-Net (ours)	72.5	54.1

Table 1. Comparison of MuCaLe-Net to baseline-strategies on two datasets. For random methods (RD and RG), we conducted four random draws and report the mean and standard-deviation.

Set of Label-Sets ($\mathcal{B}$)	VOC07	NWO
$\{\mathcal{C}\}$ (Subordinate)	70.3	51.2
$\{\mathcal{L}_1\}$ (Basic-Level)	70.0	51.0
$\{\mathcal{L}_2\}$ (Superordinate)	58.9	37.4
$\{\mathcal{C},\mathcal{L}_1\}$	72.5	54.1
$\{\mathcal{C},\mathcal{L}_1,\mathcal{L}_2\}$	**73.0**	**54.9**

Table 2. Impact of the set of label-sets ($\mathcal{B}$) used in MuCaLe-Net.

Accuracy (Acc.) for mono-label benchmarks. The cost parameter of the SVM is optimized for each dataset through cross-validation on the usual train/val splits.

Implementation details. As depicted in Sec. 3.2, we set $\mathcal{B} \in \{\mathcal{C},\mathcal{L}_1,\mathcal{L}_1\}$ with $\mathcal{C}$, $\mathcal{L}_1$ and $\mathcal{L}_2$ respectively corresponding to the *subordinate*, *basic* and *superordinate* label-sets. Practically, we used the 483 *subordinate* categories released in [5] for $\mathcal{C}$, 200 *basic* ones released in [30] for $\mathcal{L}_1$ and we re-labeled $\mathcal{L}_1$ according to our re-labeling protocol (detailed in supp. material) to get 12 *superordinate* categories for $\mathcal{L}_2$. We used the WordNet hierarchy as input for our re-labeling function (Eq. (2)). We respectively set K to 7 (15) in Eq. (5) for AlexNet (VGG). Thus, we always extract the penultimate layer ($fc7$) from each categorical-level network. Regarding the normalization step, the infinite-norm (L_∞) obtains slightly better results than L_1 or L_2-norms, thus L_∞ is applied in the following.

5.2. Comparison with Diversification-Strategies

In this section, we compare the diversification ability of the proposed MuCaLe-Net strategy with baselines and strategies of the literature. The diversification ability is defined as the ability of a strategy to produce an image representation that is more universal (more performing in a transfer-learning scheme), than a representation obtained from a reference strategy. The comparison is thus carried on a transfer-learning scheme through four target-datasets. For all the strategies, we used the same database (ILSVRC$^{0.5}$) to train the networks (AlexNet, if not specified) and the same transfer-learning protocol.

Comparison with Baseline-Strategies
We compare out method to three baseline-methods, namely **RD** (Random Difference), **RG** (Random Grouping) and **JT** (Joint-Training). RD and RG are ensemble-model

Strategy	Add. Im.	VOC07 mAP	VOC12 mAP	CA101 Acc.	NWO mAP
Standard	n/a	70.3	70.6	79.6	51.2
NSD	$+10^5$	71.7	71.7	82.6	52.9
NGD	$+10^5$	72.1	72.2	82.2	53.1
NNILM	$+10^8$	64.7*	n/a	n/a	n/a
NSM	$+10^5$	69.9	70.0	80.9	49.0
Ours	**+0**	**72.5**	**72.2**	**82.6**	**54.1**
Standard	n/a	79.9	79.2	83.5	59.3
NGD	$+10^5$	81.7	81.1	85.1	62.1
Ours	**+0**	**83.0**	**82.6**	**88.5**	**65.2**

Table 3. Comparison of diversification-strategies on four datasets with AlexNet and VGG (three last lines) networks. For VGG, we compare our method only with the best strategy (NGD). To be fair, we only used two models ($\mathcal{B} \in \{\mathcal{C},\mathcal{L}_1\}$) in our method. In the second column, we highlight the number of unique images and annotations added compared to the standard strategy. Evaluation metrics are specified under each dataset name. Results marked with * are those reported in the original papers.

methods containing the Standard network (details in next subsection) and another one: for RD, it is a network trained on the same specific database than the former network but with another (random) initialization of the weights; for RG, it is one trained on the re-labeled database obtained by grouping its specific categories into G subsets through random partitioning, instead of semantic partitioning (our method). JT consists to jointly train *one* network using both annotations (specific and generic) through a multi-label loss layer (hinge-loss). It can be seen as a multi-task learning strategy. The results are presented in Table 1 in which our method outperforms all the baselines. More specifically, the performances of RD are very close to the Standard method meaning that the random initialization of the weights does not relevantly diversify the features like our method. Even with RG, the results are also very close to the Standard method (and thus far from our method), which clearly highlights the utility of the *semantic* partitioning of our method. Another salient result is that, JT increases the performances compared to the Standard strategy, but the improvement is below our proposal, which clearly demonstrates the utility of the disjoint-learning (ours) compared to the joint one (JT). To have fair comparisons in this experiment, we only used two networks in our method (*i.e*, MuCaLe-Net with $\mathcal{B} \in \{\mathcal{C},\mathcal{L}_1\}$), but others could be used. Hence, we evaluated the impact of the set of label-sets used in our method (results in Table 2).

Comparison with Strategies of the Literature
The reference strategy and the strategies of the literature used for comparison are: **Standard** [18, 32]: Training a CNN on the 483 specific categories of the training-database. It corresponds to the standard learning strategy, thus we use it as reference for all the diversification-

Method	Network	Caltech-101 Accuracy	Caltech-256 Accuracy	VOC 2007 mAP	VOC 2012 mAP
Krizhevsky *et al.* [18]*	AlexNet	87.8	70.8	76.1	75.9
Chatfield *et al.* [3]	VGG-S	87.8	77.6	79.7	82.9
Szegedy *et al.* [33]*	GoogleNet	90.5	77.7	82.7	81.9
Simonyan *et al.* [32]*	VGG-16	88.8	78.0	86.1	84.5
He *et al.* [13]*	ResNet-50	90.8	78.9	84.4	83.1
He *et al.* [13]*	ResNet-101	91.4	80.1	85.6	84.4
Our approach	AlexNet	89.4	73.4	77.5	77.2
	VGG-16	**92.0**	**80.9**	**87.5**	**86.1**

Table 4. Comparison of the proposed MuCaLe representation with the state-of-the-art CNN-based descriptors on an image classification context over four publicly available benchmarks. Results of the methods marked with * are obtained using online code.

strategies. **NSD** [2, 41] (New Specific Data): Adding 100 new fine-grained classes (randomly obtained from ImageNet) and thus $100,000$ new images manually annotated, resulting on the learning of a network on 533 specific categories. **NGD** [18, 21] (New Generic Data): Adding 100 new generic categories and thus $100,000$ new images manually annotated, resulting to the joint learning of 483 specific categories and 100 generic ones. The generic categories are obtained from the internal-nodes of the ImageNet hierarchy at random levels. **NNILM** [17] (New Noisy Images and Labels Model): Concatenation of the features extracted from two CNN-models, the Standard one and one trained on 100 million noisy labeled images collected from Flickr. For hard-reproducibility reasons, we present the results they reported in the paper. **NSM** [2] (New Specific Model): Concatenation of the features extracted from two CNN-models, the standard one and one trained on $100,000$ images annotated among 100 new categories (randomly obtained from ImageNet).

Results are presented in Table 3. We observe that the MuCaLe-Net strategy always performs better than the standard one, regardless the network architecture (+2.4 absolute points for AlexNet and +4.3 for VGG). Moreover, MuCaLe-Net outperforms all other state-of-the-art techniques with a big advantage to be at zero-cost additional data and near-zero cost annotations. This experiment clearly demonstrates the diversification ability of the proposed learning-strategy and confirms the analyses conducted in previous sections.

5.3. Comparison with the State-of-the-Art

In this section, we compare the performances of the MuCaLe representation with state-of-the-art CNN-based descriptors on four publicly available benchmarks of image classification in a transfer-learning scheme. To get better performances, many methods [7, 24, 37] assist the classification pipeline with costly object-detectors (by extracting a lot of regions per image). Such a costly refinement is out of the scope of the paper. Hence, the comparison is carried with methods that use only the full image on the target-

datasets and we report the results obtained with online code of the best comparison methods. Regarding our method, we implemented it with the whole ILSVRC dataset and two network architectures, namely AlexNet and VGG-16. More implementation details are given in supp. material.

Comparison results are reported in Table 4, in which we observe that our proposal outperforms all state-of-the-art methods, on the four benchmarks. More specifically, our approach based on AlexNet is always better than the standard one [18]. Compared to [32], the proposed MuCaLe descriptor learned with the same network architecture (VGG-16), significantly increases their performances with an absolute improvement of 2.8% on Caltech-101, 2.9% on Caltech-256, 1.5% on VOC07 and 1.6% on VOC12. Moreover, it is worth noting that in [32], the network has been learned with scale-jittering, *i.e*, five scales of the training-images, while here we only use one scale (224×224). Another salient result is that our approach that is based on a network of 16 layers (VGG-16) desirably outperforms (on the four benchmarks) the approach of He *et al.* [13] that has been learned with networks of 50 and even 101 layers. These results clearly illustrate the positive impact of diversifying the features (*e.g.*, incorporating new relevant convolutional and unit-filters in the features' space) generated by deep CNNs through categorical-levels variation.

6. Conclusion

We proposed a novel ensemble-model strategy for learning CNNs that consists to vary the discriminative problems solved by each network of the ensemble. The variation of the discriminative problems is based on *categorical-levels* used in the human-categorization process. We have shown good image classification results on four publicly available benchmarks in a transfer-learning scheme. We also shown, through an in-depth analysis, that our method works better because of its two core properties, namely "filters-diversification" and "diversification-relevance". The findings here are quite promising and we will pursue the work by investigating whether other kind of discriminative problems variation could desirably respect the same properties.

References

[1] K. Ahmed, M. H. Baig, and L. Torresani. Network of experts for large-scale image categorization. In *ECCV*, 2016. 2

[2] H. Azizpour, A. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Factors of transferability for a generic convnet representation. *PAMI*, 2015. 1, 2, 8

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 8

[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 *CIVR*, 2009. 6

[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. 3, 4, 7

[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 *CVPR*, 2012. 2, 3

[7] T. Durand, N. Thome, and M. Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *CVPR*, 2016. 8

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. *IJCV*, 2010. 6

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

[10] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 2006. 6

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 5

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 6

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[14] L. Herranz, S. Jiang, and X. Li. Scene recognition with cnns: objects, scales and dataset bias. In *CVPR*, 2016. 1, 2, 6

[15] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 1, 2

[16] P. Jolicoeur, M. A. Gluck, and S. M. Kosslyn. Pictures and names: Making the connection. *Cognitive Psychology*, 1984. 2, 3

[17] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016. 1, 2, 8

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 4, 6, 7, 8

[19] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *ICLR*, 2016. 5

[20] A. Mathews, L. Xie, and X. He. Choosing basic-level concept names using visual and language context. In *WACV*, 2015. 2

[21] P. Mettes, D. Koelma, and C. G. M. Snoek. The imagenet shuffle: Reorganized pre-training for video event detection. In *ICMR*, 2016. 1, 2, 8

[22] V. N. Murthy, V. Singh, T. Chen, R. Manmatha, and D. Comaniciu. Deep decision network for multi-class image classification. In *CVPR*, 2016. 2

[23] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 1

[24] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 8

[25] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. Berg. From large scale image categorization to entry-level categories. In *ICCV*, 2013. 2

[26] V. Ordonez, W. Liu, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. Predicting entry-level categories. *IJCV*, 2015. 2

[27] W. Ouyang, X. Wang, C. Zhang, and X. Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *CVPR*, 2016. 2

[28] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *PAMI*, 2005. 6

[29] E. Rosch. Principles of categorization. *Cognition and Categorization*, 1978. 2, 3

[30] 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. *IJCV*, 2015. 6, 7

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014. 1, 2

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6, 7, 8

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 8

[34] Y. Tamaazousti, H. Le Borgne, and C. Hudelot. Diverse concept-level features for multi-object classification. In *ICMR*, 2016. 2

[35] J. W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 1991. 3

[36] Y. Wu, J. Li, Y. Kong, and Y. Fu. Deep convolutional neural network with independent softmax for large scale face recognition. In *ACM*, 2016. 1, 2

[37] H. Yang, J. T. Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai. Exploit bounding box annotations for multi-label object recognition. In *CVPR*, 2016. 8

[38] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014. 1, 5

[39] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. In *ICML*, 2015. 5

[40] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1, 2, 5

[41] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 1, 2, 8