Learning and Transferring Deep ConvNet Representations with Group-Sparse Factorization

Liangke Gui and Louis-Philippe Morency
School of Computer Sciences, Carnegie Mellon University
{liangkeg, morency}@cs.cmu.edu

Abstract

Deep convolutional neural networks (Deep ConvNets or CNNs) have exhibited their promise as a universal image representation for recognition. In this work we explore how the transferability of such deep ConvNet representations trained on large-scale annotated object-centric datasets (ImageNet) can be further enhanced for other visual recognition tasks with limited amount of unlabeled training data. We use group-sparse non-negative matrix factorization (GSNMF), a variant of NMF, to identify a rich set of high-level latent variables built from the pre-trained Imagenet deep ConvNets that are informative across scene and fine-grained recognition tasks. The resulting architecture can itself be seen as a feed-forward model that combines deep ConvNets and two-layer structured NMF. We demonstrate state-of-the-art image clustering performance on challenging scene (MIT-67) and fine-grained (Birds-200, Flowers-102) benchmarks. The consistent superior performance of our GSNMF-CNN shows that it is more generic for novel tasks/categories compared to the deep ConvNets activations.

1. Introduction

Deep convolutional neural networks (Deep ConvNets) [17] have recently demonstrated breakthrough performance on various visual recognition and image processing tasks [15, 29, 10]. Compared to hand-crafted or shallow features, an attractive property of deep ConvNets is their transferability. Once trained on a large corpus of annotated source data, deep ConvNets tend to learn powerful generic image representations, which can be either used off-the-shelf on the target tasks or through fine-tuning when enough labeled training data is available [1, 26].

Despite such universal and invariant representation across different tasks, the transferability of ConvNets is still limited [20]. Discrepancy of different domains still could not be removed. Dataset shift as a bottleneck to the transferability of ConvNets results in statistically unbounded risk for target tasks [2, 33]. As found in [1], the transferability is clearly correlated with the distance of the target task from the source task. In particular, in situations with limited amount of *unlabeled* training data, where conventional fine-tuning with Back-propagation is infeasible, *e.g.*, image

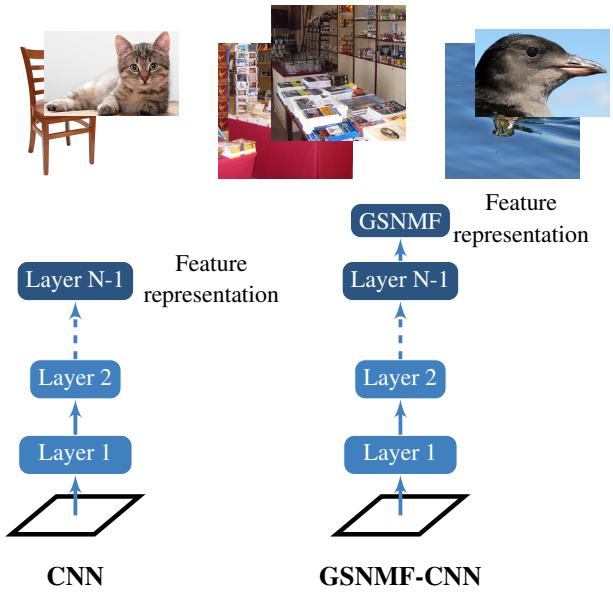


Figure 1: Transfer Deep ConvNet Representation via Group-Sparse Non-Negative Matrix Factorization. ConvNets learn invariant factors underlying different populations (**left**). Pre-trained on ImageNet, off-the-shelf deep ConvNet features are limited to describe complex indoor scenes and subtle differences among fine-grained categories. We feed the pre-trained deep ConvNet representations to an additional NMF layer with group-sparse regularization as a unified feed-forward model (**right**). We learn a better representation with enhanced transferability for target tasks, especially with limited unlabeled training data where conventional fine-tuning with Back-propagation is infeasible.

clustering, how to enhance the transferability of deep ConvNet representations is still an open challenge.

In the unsupervised scenarios, it is known that non-negative matrix factorization (NMF) and its variants are able to disentangle exploratory factors of variations underlying data samples, and have been successfully applied in many applications such as image processing[34], clustering and classification [35, 32]. Similar to ConvNets, NMF is also based on certain physiological and psychological evidence that perception of the whole is based on perception of its parts [30]. By incorporating the non-negativity constraints into the linear decomposition model, NMF obtains parts-based representations and thus enhances the corresponding interpretability.

Meanwhile, the deep ConvNet activations of interest are those after the rectified linear units (ReLUs), which consistently show better recognition performance for various tasks and which are also non-negative. Hence, it is a natural way to combine both deep ConvNets and NMF as shown in Fig. 2, which could potentially learn more semantic and meaning components on the target tasks, leading to enhanced transferability. More precisely, to select a group of correlated deep ConvNet activations, we introduce a variant of NMF—group-sparse non-negative matrix factorization (GSNMF), to identify a rich set of informative and discriminative latent variables across tasks. Given that NMF could also be interpreted as a two-layer neural network [18], our GSNMF-CNN model can be regarded as a principled feed-forward model.

Our main contribution is thus three-fold: First, we show how such a new network GSNMF-CNN, based on combining non-negative matrix factorization and pre-trained ConvNet can be operationalized for learning a more generic feature representation across tasks and datasets (Section 3). Second, we explicitly enforce group-sparsity on GSNMF-CNN to better leverage the correlation of deep ConvNet activations by introducing elastic net regularization into NMF (Section 3). Third, we show how our representation can be applied in unsupervised image clustering tasks. To the best of our knowledge, we are the first to evaluate the performance of image clustering on challenging large-scale scene and fine-grained recognition datasets (Section 4).

2. Related Work

This section reviews related work, including Deep ConvNets, NMF, and image clustering. This section also introduces notations and formulae used throughout this paper.

2.1. Deep ConvNets

For an image $\mathcal{I}$, convolutional neural networks (ConvNets or CNNs) learn a nonlinear representation $\mathcal{I} \rightarrow \Phi(\mathcal{I}) = X$ from a large corpus of annotated data in a hierarchical fashion in accordance with their relatedness to invariant factors [3]. *AlexNet* [16], the pioneering ConvNet pre-trained with millions of images from the ImageNet dataset [9], has 5 convolutional layers and 3 fully-connected layers. This model shows a good generalization ability and transferability. A number of attempts have been made to improve the original architecture to achieve better recognition accuracy. To be consistent with recent work, we use the state-of-the-art VGG ConvNet [28], which uses very small 3×3 receptive fields, but with more layers—16 convolutional layers and 3 fully-connected layers. This model produces superior performance in ILSVRC classification challenge [27] and demonstrates better transferability on other tasks and datasets.

2.2. Non-negative Matrix Factorization

Given an *M*-D random vector X with non-negative elements, *e.g.* the deep ConvNet activations, whose *N* observations are denoted as $x_j, (j = 1, 2, ..., N)$, let data matrix be $\mathbf{X}$=$[x_1, x_2, ..., x_N]$. NMF seeks non-negative basis matrix $\mathbf{W} \in \mathbb{R}^{M \times L}_{\geq 0}$ and coefficient martrix $\mathbf{H} \in \mathbb{R}^{L \times N}_{\geq 0}$, such that

$$\mathbf{X} \approx \mathbf{WH}. \tag{1}$$

Usually *L* satisfies $L \ll min(M,N)$. One commonly used object function is squared Euclidean distance

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \parallel \mathbf{X} - \mathbf{WH} \parallel^2_F, s.t. \mathbf{W}, \mathbf{H} \geq 0. \tag{2}$$

The optimization of NMF is non-convex and can be solved by alternating minimization

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} * \frac{\mathbf{XH^T}}{\mathbf{WHH^T}} \\ \mathbf{H} \leftarrow \mathbf{H} * \frac{\mathbf{W^T X}}{\mathbf{W^T WH}}. \end{cases} \tag{3}$$

Here the coefficient matrix H is new learnt feature representation.

2.3. Image Clustering

Image clustering is a challenging problem in computer vision and image processing, due to the great diversity of image contents and numerous variations in illumination and scale conditions. Fergus et al. [13] modeled objects as constellations of visual parts and used EM algorithm for unsupervised recognition parameter estimation. These methods assume that the samples have explicit distributions. However, images are arranged in complex and widely diverging shapes, making these models difficult to transfer other image datasets. The majority work uses hand-crafted features such as SIFT [21], HOG [8] and is evaluated on small datasets like or subset Image classes. The applications of deep ConvNet features to image clustering and evaluation based on large scale datasets still remain unexplored.

3. Our Approach

ConvNets can learn representations that are transferable across different tasks. However, the domain discrepancy remains especially if there is a huge statistical difference among various vision tasks. For a new unlabeled target dataset, this assumption may result in statistically unbounded risk. Meanwhile, due to the non-negativity and purely additive property, NMF could learn latent part-components which are physically meaningful in many kinds of real-world data. Thus we combine NMF and ConvNets to form a feed-forward model that can identify meaning components from non-negative representations learnt by ConvNets.

3.1. GSNMF-CNN Model

As off-the-shelf deep ConvNet features are global representations of the whole image, the generalization ability and transferability are weakened. It has been showed that Joint l_1 and l_2 penalties enjoy a similar sparsity of representation as l_1 norm and encourages a grouping effect as l_2 norm [19]. Thus we impose a weighted mixture of l_1 and squared l_2 penalities on coefficient matrix $\mathbf{H}$ to achieve group-sparse representations. The objective function is defined as (4):

$$f(\mathbf{W},\mathbf{H}) = \frac{1}{2} \parallel \mathbf{X} - \mathbf{WH} \parallel_F^2 + \frac{\alpha}{2} \parallel \mathbf{H} \parallel_2^2 + \beta \parallel \mathbf{H} \parallel_1, \quad s.t. \mathbf{W},\mathbf{H} \geq 0. \tag{4}$$

3.2. Optimization

Since we impose group-sparsity on the coefficient matrix $\mathbf{H}$, the updating rule for $\mathbf{W}$ is the same as the standard NMF in (3). To minimize equation (4), a gradient-descent based method is used and the first-order update rule of $\mathbf{H}$ should be generally

$$\mathbf{H} \leftarrow \mathbf{H} - \eta * \frac{\partial \mathbf{f}(\mathbf{H})}{\partial \mathbf{H}}, \tag{5}$$

where matrix η is the step. We take the derivative of $f(\mathbf{H})$ in Equation (4) with respect to $\mathbf{H}$

$$\frac{\partial f}{\partial \mathbf{H}} = -\mathbf{W^T X} + \mathbf{W^T WH} + \alpha \mathbf{H} + \beta. \tag{6}$$

Here we let the adaptive step size η to be

$$\eta = \frac{\mathbf{H}}{\mathbf{W^T WH} + \alpha \mathbf{H} + \beta}. \tag{7}$$

Then we get the updating rule

$$\mathbf{H} \leftarrow \mathbf{H} * \frac{\mathbf{W^T X}}{\mathbf{W^T WH} + \alpha \mathbf{H} + \beta}. \tag{8}$$

Here the coefficient $\mathbf{H}$ is the new feature representation.

4. Experimental Evaluation

In this section, we evaluate the transferability of our GSNMF-CNN representations on multiple challenging benchmarks for image clustering, where no labeled data is provided. We first introduce the datasets and the implementation details, and then present quantitative and qualitative results by comparing several state-of-the-art methods and validating across tasks the generality of GSNMF-CNN. In absolute terms, we achieve the best performance ever reported on all these benchmarks for image clustering by a significant margin.

4.1. Tasks and Datasets

Current image clustering algorithms are usually evaluated under small-scale experimental setups [7], by either using relatively simple datasets (*e.g.*, COIL-20 [23]), or sampling a portion of categories from a large dataset (*e.g.*, using 4, 7 and 20 sub-categories of the Caltech-101 dataset [11]). Unlike previous work and consistent with [7], to show the transferability of our GSNMF-CNN representations across categories and tasks, we consider using them for *unsupervised scene and fine-grained image clustering* on large-scale benchmarks including the MIT-67 [25], Caltech-UCSD Birds (CUB) 200-2011 [36] and Oxford 102 flowers [24] datasets.

These are very challenging tasks since 1) There are strong domain shifts between the source and target datasets. Compared to the object-centric ILSVRC dataset where the deep ConvNet features are pre-trained, the target MIT-67 dataset is more scene-centric and consists of similar objects presented in different indoor scenes [25], and the target Birds-200 and Flowers-102 datasets involve very subtle differences between examples of a visual category [1]. Importantly, the transferability of a ConvNet decreases when the target task is far from the ConvNet source task [1], as in our case. 2) The datasets used for evaluation are standard classification benchmarks, and they are still very challenging even for supervised image classification. However, we tackle a more difficult scenario here by testing the representations for unsupervised image clustering, without having access to the label information on these datasets. We will show that with limited amount of unlabeled training data from distinct target tasks, our GSNMF-CNN model is capable of discovering informative and discriminative latent variables from deep ConvNet activations.

We follow the standard experimental setup (*e.g.*, the train/test splits) for these dataset during our experiments. A brief description of the datasets is as follow:

- MIT-67 [25]: MIT-67 consists of 15K image spanning 67 indoor scene classes such as shoe shop, mall and garage. As it has a significant statistics difference from the ImageNet, indoor scenes tend to vary a lot in term

of composition and better characterized by the objects they contain. This makes it more challenging and an interesting test case for the feature representation. The provided training/testing split for this dataset consist of 80 training and 20 testing images per class.

- Caltech-UCSD Birds (CUB) 200-2011 dataset [36]: Birds-200 contains 11788 images of 200 birds species (mostly North American). 5994 images are used for training and 5794 for testing. As a fine-grained recognition dataset, many of the species in the dataset exhibit extremely subtle differences which are sometimes even hard for humans to distinguish. Bird bounding boxes, 15 part landmarks, 312 binary attributes and boundary segmentation are available for this dataset. In this work we only use the bounding box annotation during training and testing.

- Oxford 102 Flowers [24]: it contains 102 flower categories and each class consists of between 40 and 258 images and 10 images are used as training data and the rest as testing data. Additionally, the dataset provides segmentation for all the images. The subtlety of difference across different subclasses require a fine-detailed feature representation which makes fin-grained recognition a good test of whether a generic representation can capture these subtle details.

4.2. Implementation Details

Our feed-forward GSNMF-CNN model includes two modules. For the deep ConvNet layers, we use the Caffe VGGNet pretrained on ILSVRC (All the weights of the deep ConvNet are frozen to those learned on ILSVRC without fine-tuning on any other datasets) [14, 28]. For each image, we extract features on the center 224×224 crop of the 256×256 resized image. It is a $d=4{,}096$-D feature vector $fc7$ taken from the last hidden layer of the network. For the GSNMF layers, we use the unlabeled training data on the target task to learn the weights or bases. L is set to be $1{,}024$. The test images are then fed forward to the learned GSNMF-CNN model, producing a final d=1,024-D feature representation. As our main purpose is to validate whether the proposed approach is able to boost the transferability of the deep ConvNet features for image clustering, we use the standard clustering algorithm—*k*-means for fair comparison. All experiments were done following the standard training/test splits.

Parameter settings. For the parameters α and β in GSNMF, in a preliminary experiment, we tested image clustering on the Scene15 dataset [12], which is relatively small dataset. After searching α and β on a 2D grid $10^{[-3:1:2]} \times 10^{[-3:1:2]}$, we observed that the best performance was achieved when $\alpha = 10$ and $\beta = 10$. In all our experiments, we then simply set α and β to be 10. Even better performance could be obtained by further tuning them.

Baselines. We compare our GSNMF-CNN feature against the following related representations: 1) deep ConvNet, the original pre-trained deep ConvNet feature of size $4{,}096$; 2) PCA-CNN, which applies principal components analysis (PCA) to deep ConvNets, leading to $d=1{,}024$-D feature vectors; 3) EP-CNN, which generates an ensemble of classifiers based on deep ConvNet features and represents images by the concatenation of their classification scores [5, 6, 7]. We use the same setup and default parameters as in [7], leading to $d=3{,}000$-D feature vectors. All the representation models, such as the PCA bases and the ensemble classifiers, are learned on the unlabeled training dataset, and their discriminative ability is evaluated on the test dataset for *k*-means clustering. Note that in our clustering scenarios, without labeled data, we cannot conduct conventional fine-tuning with Backpropagation on the target dataset.

Evaluation Metrics: To be consistent with previous work, Accuracy and Normalized Mutual Information [4, 37] are used as the evaluation criterion. For the definition of accuracy, suppose the clustering algorithm is tested on *H* samples. For a sample x_i, the cluster label is denoted as r_i, and ground truth is t_i. The accuracy is defined as follows:

$$accuracy = \frac{\sum_{i=1}^{N} \delta(t_i, map(r_i))}{N}, \tag{9}$$

where $\delta(x, y)$ equals to 1 if x is equal to y and 0 otherwise. Function $map(x)$ denotes the best permutation mapping function gained by Kuhn-Munkres algorithm, which maps cluster to the corresponding predicted label. So the more labels of samples are predicted correctly, the greater the accuracy is.

For Normalized Mutual Information (NMI), let C denotes the cluster centers of ground truth, and $C^{'}$ denotes the cluster centers by clustering algorithm. The NMI is defined as follows:

$$NMI(C, C^{'}) = \frac{MI(C, C^{'})}{max(H(C), H(C^{'})}, \tag{10}$$

where $H(C)$ and $H(C^{'})$ are the entropies of C and $C^{'}$. $MI(C, C^{'})$ is the mutual information of C and $C^{'}$. NMI measures the dependency of two distributions and higher value of NMI means greater similarity between two distributions.

4.3. Image Clustering Results

Table 1 and Table 2 summarize the *K*-Means clustering performance of our GSNMF-CNN representation and related baseline features on the scene and fine-grained recognition datasets. We can see that our GSNMF-CNN representation outperforms the original CNN feature and its PCA

	MIT-67	Birds-200	Flowers-102
CNN	0.471	0.361	0.474
PCA-CNN	0.464	0.360	0.482
EP-CNN [5, 6, 7]	0.420	0.354	0.453
GSNMF-CNN	**0.509**	**0.389**	**0.528**

Table 1: Accuracy of scene and fine-grained image clustering on three large-scale benchmark datasets. Ours significantly outperform all the other baselines.

	MIT-67	Birds-200	Flowers-102
CNN	0.646	0.661	0.660
PCA-CNN	0.653	0.661	0.660
EP-CNN [5, 6, 7]	0.619	0.646	0.632
GSNMF-CNN	**0.661**	**0.670**	**0.671**

Table 2: Normalized Mutual Information of scene and fine-grained image clustering on three large-scale benchmark datasets. Ours significantly outperform all the other baselines.

transformed version by a considerable margin. For example, in terms of clustering accuracy, GSNMF-CNN outperforms CNN by 5.2% on MIT-67, 3.3% on Birds-200 and 1% on Flowers-102. Moreover, EP-CNN reported improved performance over CNN in transductive learning, where the EP representation (ensemble of classifiers) is learned using both the training and test datasets [5]; however, in our case of learning representation on the training dataset and conducting clustering on the test dataset, EP-CNN shows inferior performance to CNN. This means that having access to the distribution of the test data is advantageous for EP-CNN. The superior performance of our GSNMF-CNN reveals that it could learn a more generic and transferable representation to capture the subtlety of differences across different subordinate classes and tasks.

Size of Training Dataset. We evaluate clustering performance as a function of the number of training examples per class on the MIT-67 dataset. For the standard training/test split (80 training and 20 test images per class), we randomly select 40, 60 images out of the 80 training images per class for training, and use all the same 20 test images for testing. Fig. 2 summarizes the average performance over 10 random splits. In all cases, our GSNMF-CNN outperforms the baseline approaches, and shows consistently improved performance with more training data.

4.4. Qualitative Visualization

We visualize the model features to gain insight into the semantic capacity and transferability of GSNMF-CNN.

Semantic Groups. By using the t-SNE algorithm [31], we find a two-dimensional embedding of the high-dimensional feature space, and plot them as points colored depending on their ground-truth labels. Since it is visually difficult to represent and distinguish too many classes on the t-SNE embedding, we randomly select 10 classes of the Birds-200 dataset. As shown in Fig. 3, compared to CNN and PCA-CNN, GSNMF-CNN shows very good clustering of semantic classes, even for the fine-grained categories with extremely subtle differences. This behavior explains the improved clustering performance of GSNMF-CNN in Table 1 and Table 2.

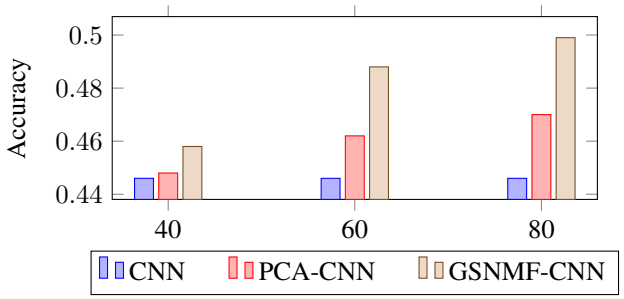


Figure 2: Clustering result based for scene clustering on the MIT-67 dataset. X-axis: number of training examples per class. Y-axis: average clustering accuracy.

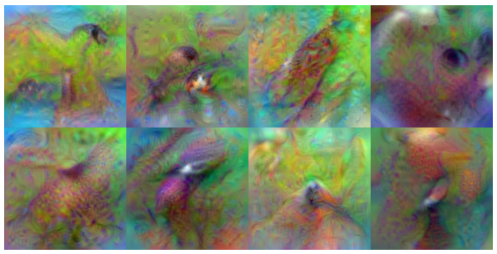
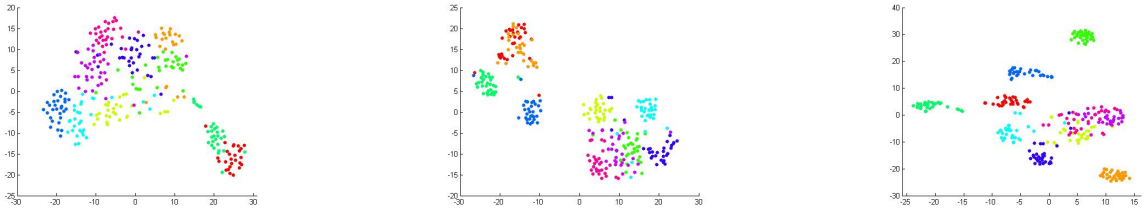


Figure 4: Representative inversion and reconstruction of GSNMF-CNN bases in the image space on the Birds-200 dataset. GSNMF successfully identifies some localized features that correspond with intuitive notions of the parts of bird species, such as eyes, beak and wings.

Bases Inversion. By using the representation inversion technique [22], we reconstruct the learned GSNMF bases in the image space. We show some representative reconstructions on the Birds-200 dataset in Fig. 4. Consistent with the conventional NMF, now in the CNN feature space, GSNMF also identifies some localized features that correspond with intuitive notions of the parts of bird species, such as eyes, beak and wings, in the image space.

5. Conclusion

In this paper, we have addressed to boost the transferability of a deep ConvNet representation for other visual recognition tasks with limited amount of *unlabeled* training



(a) CNN (b) PCA-CNN (c) GSNMF-CNN

Figure 3: t-SNE feature visualizations on 10 random classes of the Birds-200 dataset for (a) CNN, (b) PCA-CNN, and (c) GSNMF-CNN. The better clustering behavior suggests that GSNMF-CNN is a transferable representation even for fine-grained categories with subtle differences.

data, where conventional fine-tuning with Backpropagation is infeasible. By imposing group-sparse non-negative matrix factorization (GSNMF) on deep ConvNet activations to constitute a feed-forward model, we discovered a rich set of informative and discriminative latent variables. Extensive large-scale image clustering experiments confirm that the new feature representations are significantly universal for scene and fine-grained recognition tasks.

References

[1] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. From generic to specific deep representations for visual recognition. *arXiv:1406.5774*, 2014.

[2] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al. Analysis of representations for domain adaptation. *NIPS*, 2007.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *TPAMI*, 35(8):1798–1828, 2013.

[4] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *TPAMI*, 33(8):1548–1560, 2011.

[5] D. Dai, M. Prasad, C. Leistner, and L. Van Gool. Ensemble partitioning for unsupervised image categorization. In *ECCV*. 2012.

[6] D. Dai and L. Van Gool. Ensemble projection for semi-supervised image classification. In *ICCV*, 2013.

[7] D. Dai and L. Van Gool. Unsupervised high-level feature learning by ensemble projection for semi-supervised image classification and image clustering. Technical report, ETH Zurich, May 2015.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[9] 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.

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *ICML*, 2014.

[11] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, 2007.

[12] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[13] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

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

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[17] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a backpropagation network. In *NIPS*, 1990.

[18] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[19] W. Liu, S. Zheng, S. Jia, L. Shen, and X. Fu. Sparse nonnegative matrix factorization with the elastic net. In *BIM*, 2010.

[20] M. Long and J. Wang. Learning transferable features with deep adaptation networks. *arXiv:1502.02791*, 2015.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[22] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. *CVPR*, 2015.

[23] S. A. Nene, S. K. Nayar, H. Murase, et al. Columbia object image library (coil-20). Technical report, 1996.

[24] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICCVGIP*, Dec 2008.

[25] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPRW,2014*, pages 512–519, 2014.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 1–42, 2014.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

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

[30] S. Ullman and G. W. Humphreys. *High-level vision: Object recognition and visual cognition*, volume 2. 1996.

[31] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *JLMR*, 9(2579-2605):85, 2008.

[32] Y.-X. Wang, L.-Y. Gui, and Y.-J. Zhang. Neighborhood preserving non-negative tensor factorization for image representation. In *ICASSP*, 2012.

[33] Y.-X. Wang and M. Hebert. Model recommendation: Generating object detectors from few samples. In *CVPR*, 2015.

[34] Y.-X. Wang and Y.-J. Zhang. Image inpainting via weighted sparse non-negative matrix factorization. In *ICIP*, 2011.

[35] Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *TKDE*, 25(6):1336–1353, 2013.

[36] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[37] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *ACM SIGIR*, 2003.