

Deep Clustering for Unsupervised Learning of Visual Features

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze

Facebook AI Research

{mathilde,bojanowski,ajoulin,matthijs}@fb.com

Abstract. Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large-scale datasets. In this work, we present DeepCluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. DeepCluster iteratively groups the features with a standard clustering algorithm, k -means, and uses the subsequent assignments as supervision to update the weights of the network. We apply DeepCluster to the unsupervised training of convolutional neural networks on large datasets like ImageNet and YFCC100M. The resulting model outperforms the current state of the art by a significant margin on all the standard benchmarks.

Keywords: unsupervised learning, clustering

1 Introduction

Pre-trained convolutional neural networks, or convnets, have become the building blocks in most computer vision applications [8, 9, 50, 65]. They produce excellent general-purpose features that can be used to improve the generalization of models learned on a limited amount of data [53]. The existence of ImageNet [12], a large fully-supervised dataset, has been fueling advances in pre-training of convnets. However, Stock and Cisse [57] have recently presented empirical evidence that the performance of state-of-the-art classifiers on ImageNet is largely underestimated, and little error is left unresolved. This explains in part why the performance has been saturating despite the numerous novel architectures proposed in recent years [9, 21, 23]. As a matter of fact, ImageNet is relatively small by today’s standards; it “only” contains a million images that cover the specific domain of object classification. A natural way to move forward is to build a bigger and more diverse dataset, potentially consisting of billions of images. This, in turn, would require a tremendous amount of manual annotations, despite the expert knowledge in crowdsourcing accumulated by the community over the years [30]. Replacing labels by raw metadata leads to biases in the visual representations with unpredictable consequences [41]. This calls for methods that can be trained on internet-scale datasets with no supervision.

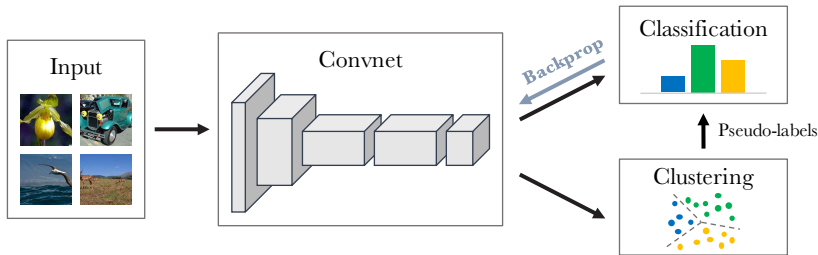


Fig. 1: Illustration of the proposed method: we iteratively cluster deep features and use the cluster assignments as pseudo-labels to learn the parameters of the convnet

Unsupervised learning has been widely studied in the Machine Learning community [19], and algorithms for clustering, dimensionality reduction or density estimation are regularly used in computer vision applications [27, 54, 60]. For example, the “bag of features” model uses clustering on handcrafted local descriptors to produce good image-level features [11]. A key reason for their success is that they can be applied on any specific domain or dataset, like satellite or medical images, or on images captured with a new modality, like depth, where annotations are not always available in quantity. Several works have shown that it was possible to adapt unsupervised methods based on density estimation or dimensionality reduction to deep models [20, 29], leading to promising all-purpose visual features [5, 15]. Despite the primeval success of clustering approaches in image classification, very few works [3, 66, 68] have been proposed to adapt them to the end-to-end training of convnets, and never at scale. An issue is that clustering methods have been primarily designed for linear models on top of fixed features, and they scarcely work if the features have to be learned simultaneously. For example, learning a convnet with k -means would lead to a trivial solution where the features are zeroed, and the clusters are collapsed into a single entity.

In this work, we propose a novel clustering approach for the large scale end-to-end training of convnets. We show that it is possible to obtain useful general-purpose visual features with a clustering framework. Our approach, summarized in Figure 1, consists in alternating between clustering of the image descriptors and updating the weights of the convnet by predicting the cluster assignments. For simplicity, we focus our study on k -means, but other clustering approaches can be used, like Power Iteration Clustering (PIC) [36]. The overall pipeline is sufficiently close to the standard supervised training of a convnet to reuse many common tricks [24]. Unlike self-supervised methods [13, 42, 45], clustering has the advantage of requiring little domain knowledge and no specific signal from the inputs [63, 71]. Despite its simplicity, our approach achieves significantly higher performance than previously published unsupervised methods on both ImageNet classification and transfer tasks.

Finally, we probe the robustness of our framework by modifying the experimental protocol, in particular the training set and the convnet architecture.

The resulting set of experiments extends the discussion initiated by Doersch *et al.* [13] on the impact of these choices on the performance of unsupervised methods. We demonstrate that our approach is robust to a change of architecture. Replacing an AlexNet by a VGG [55] significantly improves the quality of the features and their subsequent transfer performance. More importantly, we discuss the use of ImageNet as a training set for unsupervised models. While it helps understanding the impact of the labels on the performance of a network, ImageNet has a particular image distribution inherited from its use for a fine-grained image classification challenge: it is composed of well-balanced classes and contains a wide variety of dog breeds for example. We consider, as an alternative, random Flickr images from the YFCC100M dataset of Thomee *et al.* [58]. We show that our approach maintains state-of-the-art performance when trained on this uncured data distribution. Finally, current benchmarks focus on the capability of unsupervised convnets to capture class-level information. We propose to also evaluate them on image retrieval benchmarks to measure their capability to capture instance-level information.

In this paper, we make the following contributions: **(i)** a novel unsupervised method for the end-to-end learning of convnets that works with any standard clustering algorithm, like k -means, and requires minimal additional steps; **(ii)** state-of-the-art performance on many standard transfer tasks used in unsupervised learning; **(iii)** performance above the previous state of the art when trained on an uncured image distribution; **(iv)** a discussion about the current evaluation protocol in unsupervised feature learning.

2 Related Work

Unsupervised learning of features. Several approaches related to our work learn deep models with no supervision. Coates and Ng [10] also use k -means to pre-train convnets, but learn each layer sequentially in a bottom-up fashion, while we do it in an end-to-end fashion. Other clustering losses [3, 16, 35, 66, 68] have been considered to jointly learn convnet features and image clusters but they have never been tested on a scale to allow a thorough study on modern convnet architectures. Of particular interest, Yang *et al.* [68] iteratively learn convnet features and clusters with a recurrent framework. Their model offers promising performance on small datasets but may be challenging to scale to the number of images required for convnets to be competitive. Closer to our work, Bojanowski and Joulin [5] learn visual features on a large dataset with a loss that attempts to preserve the information flowing through the network [37]. Their approach discriminates between images in a similar way as exemplar SVM [39], while we are simply clustering them.

Self-supervised learning. A popular form of unsupervised learning, called “self-supervised learning” [52], uses pretext tasks to replace the labels annotated by humans by “pseudo-labels” directly computed from the raw input data. For example, Doersch *et al.* [13] use the prediction of the relative position of patches

in an image as a pretext task, while Noroozi and Favaro [42] train a network to spatially rearrange shuffled patches. Another use of spatial cues is the work of Pathak *et al.* [46] where missing pixels are guessed based on their surrounding. Paulin *et al.* [47] learn patch level Convolutional Kernel Network [38] using an image retrieval setting. Others leverage the temporal signal available in videos by predicting the camera transformation between consecutive frames [1], exploiting the temporal coherence of tracked patches [63] or segmenting video based on motion [45]. Apart from spatial and temporal coherence, many other signals have been explored: image colorization [33, 71], cross-channel prediction [72], sound [44] or instance counting [43]. More recently, several strategies for combining multiple cues have been proposed [14, 64]. Contrary to our work, these approaches are domain dependent, requiring expert knowledge to carefully design a pretext task that may lead to transferable features.

Generative models. Recently, unsupervised learning has been making a lot of progress on image generation. Typically, a parametrized mapping is learned between a predefined random noise and the images, with either an autoencoder [4, 22, 29, 40, 62], a generative adversarial network (GAN) [20] or more directly with a reconstruction loss [6]. Of particular interest, the discriminator of a GAN can produce visual features, but their performance are relatively disappointing [15]. Donahue *et al.* [15] and Dumoulin *et al.* [17] have shown that adding an encoder to a GAN produces visual features that are much more competitive.

3 Method

After a short introduction to the supervised learning of convnets, we describe our unsupervised approach as well as the specificities of its optimization.

3.1 Preliminaries

Modern approaches to computer vision, based on statistical learning, require good image featurization. In this context, convnets are a popular choice for mapping raw images to a vector space of fixed dimensionality. When trained on enough data, they constantly achieve the best performance on standard classification benchmarks [21, 32]. We denote by f_θ the convnet mapping, where θ is the set of corresponding parameters. We refer to the vector obtained by applying this mapping to an image as feature or representation. Given a training set $X = \{x_1, x_2, \dots, x_N\}$ of N images, we want to find a parameter θ^* such that the mapping f_{θ^*} produces good general-purpose features.

These parameters are traditionally learned with supervision, *i.e.* each image x_n is associated with a label y_n in $\{0, 1\}^k$. This label represents the image’s membership to one of k possible predefined classes. A parametrized classifier g_W predicts the correct labels on top of the features $f_\theta(x_n)$. The parameters W of the classifier and the parameter θ of the mapping are then jointly learned by

optimizing the following problem:

$$\min_{\theta, W} \frac{1}{N} \sum_{n=1}^N \ell(g_W(f_\theta(x_n)), y_n), \quad (1)$$

where ℓ is the multinomial logistic loss, also known as the negative log-softmax function. This cost function is minimized using mini-batch stochastic gradient descent [7] and backpropagation to compute the gradient [34].

3.2 Unsupervised learning by clustering

When θ is sampled from a Gaussian distribution, without any learning, f_θ does not produce good features. However the performance of such random features on standard transfer tasks, is far above the chance level. For example, a multilayer perceptron classifier on top of the last convolutional layer of a random AlexNet achieves 12% in accuracy on ImageNet while the chance is at 0.1% [42]. The good performance of random convnets is intimately tied to their convolutional structure which gives a strong prior on the input signal. The idea of this work is to exploit this weak signal to bootstrap the discriminative power of a convnet. We cluster the output of the convnet and use the subsequent cluster assignments as “pseudo-labels” to optimize Eq. (1). This deep clustering (DeepCluster) approach iteratively learns the features and groups them.

Clustering has been widely studied and many approaches have been developed for a variety of circumstances. In the absence of points of comparisons, we focus on a standard clustering algorithm, k -means. Preliminary results with other clustering algorithms indicates that this choice is not crucial. k -means takes a set of vectors as input, in our case the features $f_\theta(x_n)$ produced by the convnet, and clusters them into k distinct groups based on a geometric criterion. More precisely, it jointly learns a $d \times k$ centroid matrix C and the cluster assignments y_n of each image n by solving the following problem:

$$\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^N \min_{y_n \in \{0,1\}^k} \|f_\theta(x_n) - C y_n\|_2^2 \quad \text{such that} \quad y_n^\top \mathbf{1}_k = 1. \quad (2)$$

Solving this problem provides a set of optimal assignments $(y_n^*)_{n \leq N}$ and a centroid matrix C^* . These assignments are then used as pseudo-labels; we make no use of the centroid matrix.

Overall, DeepCluster alternates between clustering the features to produce pseudo-labels using Eq. (2) and updating the parameters of the convnet by predicting these pseudo-labels using Eq. (1). This type of alternating procedure is prone to trivial solutions; we describe how to avoid such degenerate solutions in the next section.

3.3 Avoiding trivial solutions

The existence of trivial solutions is not specific to the unsupervised training of neural networks, but to any method that jointly learns a discriminative classifier and the labels. Discriminative clustering suffers from this issue even when

applied to linear models [67]. Solutions are typically based on constraining or penalizing the minimal number of points per cluster [2, 26]. These terms are computed over the whole dataset, which is not applicable to the training of convnets on large scale datasets. In this section, we briefly describe the causes of these trivial solutions and give simple and scalable workarounds.

Empty clusters. A discriminative model learns decision boundaries between classes. An optimal decision boundary is to assign all of the inputs to a single cluster [67]. This issue is caused by the absence of mechanisms to prevent from empty clusters and arises in linear models as much as in convnets. A common trick used in feature quantization [25] consists in automatically reassigning empty clusters during the k -means optimization. More precisely, when a cluster becomes empty, we randomly select a non-empty cluster and use its centroid with a small random perturbation as the new centroid for the empty cluster. We then reassign the points belonging to the non-empty cluster to the two resulting clusters.

Trivial parametrization. If the vast majority of images is assigned to a few clusters, the parameters θ will exclusively discriminate between them. In the most dramatic scenario where all but one cluster are singleton, minimizing Eq. (1) leads to a trivial parametrization where the convnet will predict the same output regardless of the input. This issue also arises in supervised classification when the number of images per class is highly unbalanced. For example, metadata, like hashtags, exhibits a Zipf distribution, with a few labels dominating the whole distribution [28]. A strategy to circumvent this issue is to sample images based on a uniform distribution over the classes, or pseudo-labels. This is equivalent to weight the contribution of an input to the loss function in Eq. (1) by the inverse of the size of its assigned cluster.

3.4 Implementation details

Training data and convnet architectures. We train DeepCluster on the training set of ImageNet [12] (1, 281, 167 images distributed uniformly into 1, 000 classes). We discard the labels. For comparison with previous works, we use a standard AlexNet [32] architecture. It consists of five convolutional layers with 96, 256, 384, 384 and 256 filters; and of three fully connected layers. We remove the Local Response Normalization layers and use batch normalization [24]. We also consider a VGG-16 [55] architecture with batch normalization. Unsupervised methods often do not work directly on color and different strategies have been considered as alternatives [13, 42]. We apply a fixed linear transformation based on Sobel filters to remove color and increase local contrast [5, 47].

Optimization. We cluster the features of the central cropped images and train the convnet with data augmentation (random horizontal flips and crops of random sizes and aspect ratios). This enforces invariance to data augmentation which is useful for feature learning [16]. The network is trained with dropout [56],

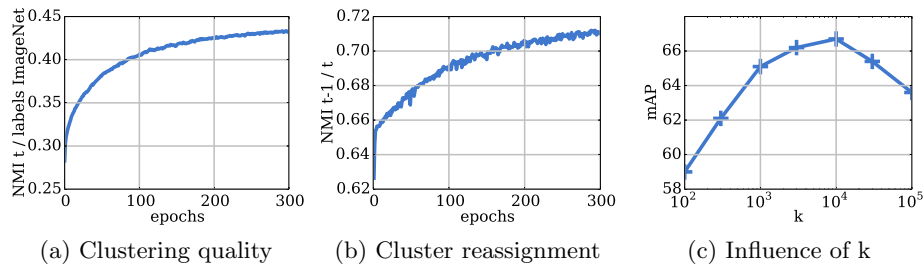


Fig. 2: Preliminary studies. (a): evolution of the clustering quality along training epochs; (b): evolution of cluster reassignments at each clustering step; (c): validation mAP classification performance for various choices of k

a constant step size, an ℓ_2 penalization of the weights θ and a momentum of 0.9. Each mini-batch contains 256 images. For the clustering, features are PCA-reduced to 256 dimensions, whitened and ℓ_2 -normalized. We use the k -means implementation of Johnson *et al.* [25]. Note that running k -means takes a third of the time because a forward pass on the full dataset is needed. One could reassign the clusters every n epochs, but we found out that our setup on ImageNet (updating the clustering every epoch) was nearly optimal. On Flickr, the concept of epoch disappears: choosing the tradeoff between the parameter updates and the cluster reassignments is more subtle. We thus kept almost the same setup as in ImageNet. We train the models for 500 epochs, which takes 12 days on a Pascal P100 GPU for AlexNet.

Hyperparameter selection. We select hyperparameters on a down-stream task, *i.e.*, object classification on the validation set of PASCAL VOC with no fine-tuning. We use the publicly available code of Krähenbühl¹.

4 Experiments

In a preliminary set of experiments, we study the behavior of DeepCluster during training. We then qualitatively assess the filters learned with DeepCluster before comparing our approach to previous state-of-the-art models on standard benchmarks.

4.1 Preliminary study

We measure the information shared between two different assignments A and B of the same data by the Normalized Mutual Information (NMI), defined as:

$$\text{NMI}(A; B) = \frac{I(A; B)}{\sqrt{H(A)H(B)}}$$

¹ <https://github.com/philkkr/voc-classification>

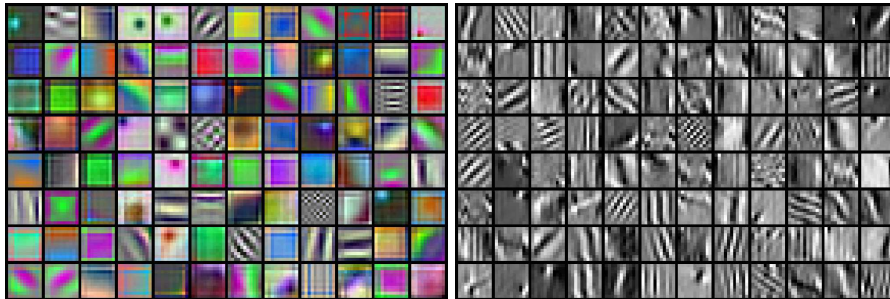


Fig. 3: Filters from the first layer of an AlexNet trained on unsupervised ImageNet on raw RGB input (left) or after a Sobel filtering (right)

where I denotes the mutual information and H the entropy. This measure can be applied to any assignment coming from the clusters or the true labels. If the two assignments A and B are independent, the NMI is equal to 0. If one of them is deterministically predictable from the other, the NMI is equal to 1.

Relation between clusters and labels. Figure 2(a) shows the evolution of the NMI between the cluster assignments and the ImageNet labels during training. It measures the capability of the model to predict class level information. Note that we only use this measure for this analysis and not in any model selection process. The dependence between the clusters and the labels increases over time, showing that our features progressively capture information related to object classes.

Number of reassignments between epochs. At each epoch, we reassign the images to a new set of clusters, with no guarantee of stability. Measuring the NMI between the clusters at epoch $t - 1$ and t gives an insight on the actual stability of our model. Figure 2(b) shows the evolution of this measure during training. The NMI is increasing, meaning that there are less and less reassignments and the clusters are stabilizing over time. However, NMI saturates below 0.8, meaning that a significant fraction of images are regularly reassigned between epochs. In practice, this has no impact on the training and the models do not diverge.

Choosing the number of clusters. We measure the impact of the number k of clusters used in k -means on the quality of the model. We report the same downstream task as in the hyperparameter selection process, *i.e.* mAP on the PASCAL VOC 2007 classification validation set. We vary k on a logarithmic scale, and report results after 300 epochs in Figure 2(c). The performance after the same number of epochs for every k may not be directly comparable, but it reflects the hyper-parameter selection process used in this work. The best performance is obtained with $k = 10,000$. Given that we train our model on ImageNet, one

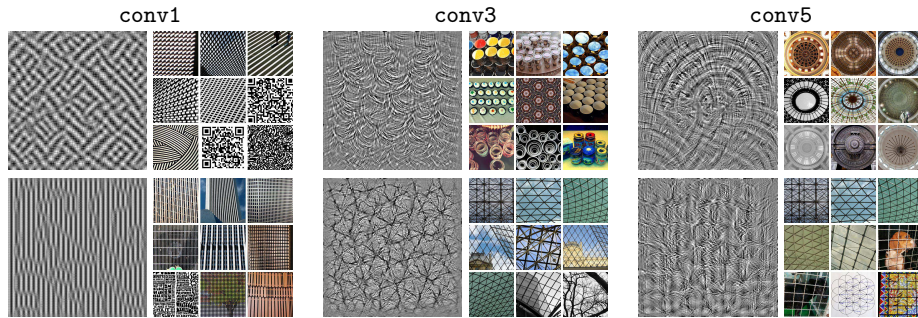


Fig. 4: Filter visualization and top 9 activated images from a subset of 1 million images from YFCC100M for target filters in the layers `conv1`, `conv3` and `conv5` of an AlexNet trained with DeepCluster on ImageNet. The filter visualization is obtained by learning an input image that maximizes the response to a target filter [69]

would expect $k = 1000$ to yield the best results, but apparently some amount of over-segmentation is beneficial.

4.2 Visualizations

First layer filters. Figure 3 shows the filters from the first layer of an AlexNet trained with DeepCluster on raw RGB images and images preprocessed with a Sobel filtering. The difficulty of learning convnets on raw images has been noted before [5, 13, 42, 47]. As shown in the left panel of Fig. 3, most filters capture only color information that typically plays a little role for object classification [61]. Filters obtained with Sobel preprocessing act like edge detectors.

Probing deeper layers. We assess the quality of a target filter by learning an input image that maximizes its activation [18, 70]. We follow the process described by Yosinki *et al.* [69] with a cross entropy function between the target filter and the other filters of the same layer. Figure 4 shows these synthetic images as well as the 9 top activated images from a subset of 1 million images from YFCC100M. As expected, deeper layers in the network seem to capture larger textural structures. However, some filters in the last convolutional layers seem to be simply replicating the texture already captured in previous layers, as shown on the second row of Fig. 5. This result corroborates the observation by Zhang *et al.* [72] that features from `conv3` or `conv4` are more discriminative than those from `conv5`.

Finally, Figure 5 shows the top 9 activated images of some `conv5` filters that seem to be semantically coherent. The filters on the top row contain information about structures that highly correlate with object classes. The filters on the bottom row seem to trigger on style, like drawings or abstract shapes.

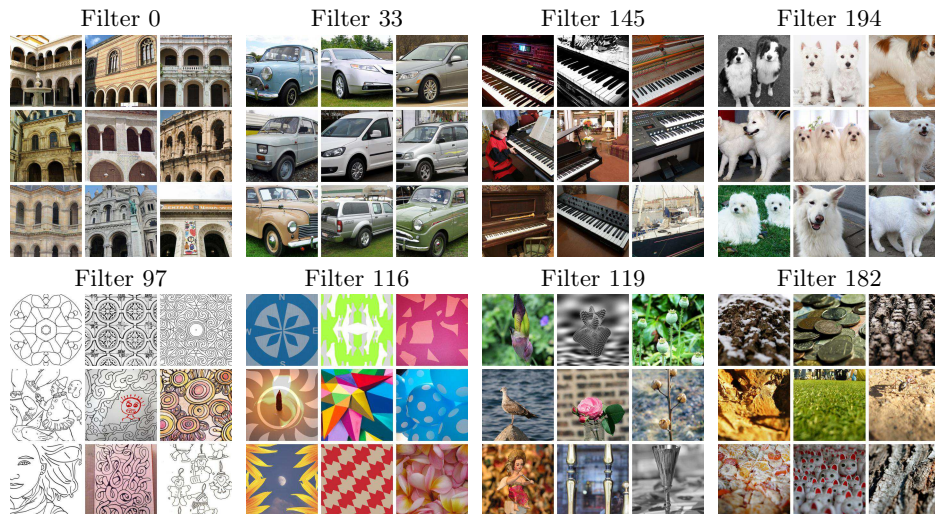


Fig. 5: Top 9 activated images from a random subset of 10 millions images from YFCC100M for target filters in the last convolutional layer. The top row corresponds to filters sensitive to activations by images containing objects. The bottom row exhibits filters more sensitive to stylistic effects. For instance, the filters 119 and 182 seem to be respectively excited by background blur and depth of field effects

4.3 Linear classification on activations

Following Zhang *et al.* [72], we train a linear classifier on top of different frozen convolutional layers. This layer by layer comparison with supervised features exhibits where a convnet starts to be task specific, *i.e.* specialized in object classification. We report the results of this experiment on ImageNet and the Places dataset [73] in Table 1. We choose the hyperparameters by cross-validation on the training set. On ImageNet, DeepCluster outperforms the state of the art from `conv2` to `conv5` layers by 1–6%. The largest improvement is observed in the `conv3` layer, while the `conv1` layer performs poorly, probably because the Sobel filtering discards color. Consistently with the filter visualizations of Sec. 4.2, `conv3` works better than `conv5`. Finally, the difference of performance between DeepCluster and a supervised AlexNet grows significantly on higher layers: at layers `conv2–conv3` the difference is only around 4%, but this difference rises to 12.3% at `conv5`, marking where the AlexNet probably stores most of the class level information. In the supplementary material, we also report the accuracy if a MLP is trained on the last layer; DeepCluster outperforms the state of the art by 8%.

The same experiment on the Places dataset provides some interesting insights: like DeepCluster, a supervised model trained on ImageNet suffers from a decrease of performance for higher layers (`conv4` versus `conv5`). Moreover,

Table 1: Linear classification on ImageNet and Places using activations from the convolutional layers of an AlexNet as features. We report classification accuracy averaged over 10 crops. Numbers for other methods are from Zhang *et al.* [72]

| Method | ImageNet | | | | | Places | | | | |
|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | conv1 | conv2 | conv3 | conv4 | conv5 | conv1 | conv2 | conv3 | conv4 | conv5 |
| Places labels | – | – | – | – | – | 22.1 | 35.1 | 40.2 | 43.3 | 44.6 |
| ImageNet labels | 19.3 | 36.3 | 44.2 | 48.3 | 50.5 | 22.7 | 34.8 | 38.4 | 39.4 | 38.7 |
| Random | 11.6 | 17.1 | 16.9 | 16.3 | 14.1 | 15.7 | 20.3 | 19.8 | 19.1 | 17.5 |
| Pathak <i>et al.</i> [46] | 14.1 | 20.7 | 21.0 | 19.8 | 15.5 | 18.2 | 23.2 | 23.4 | 21.9 | 18.4 |
| Doersch <i>et al.</i> [13] | 16.2 | 23.3 | 30.2 | 31.7 | 29.6 | 19.7 | 26.7 | 31.9 | 32.7 | 30.9 |
| Zhang <i>et al.</i> [71] | 12.5 | 24.5 | 30.4 | 31.5 | 30.3 | 16.0 | 25.7 | 29.6 | 30.3 | 29.7 |
| Donahue <i>et al.</i> [15] | 17.7 | 24.5 | 31.0 | 29.9 | 28.0 | 21.4 | 26.2 | 27.1 | 26.1 | 24.0 |
| Noroozi and Favaro [42] | 18.2 | 28.8 | 34.0 | 33.9 | 27.1 | 23.0 | 32.1 | 35.5 | 34.8 | 31.3 |
| Noroozi <i>et al.</i> [43] | 18.0 | 30.6 | 34.3 | 32.5 | 25.7 | 23.3 | 33.9 | 36.3 | 34.7 | 29.6 |
| Zhang <i>et al.</i> [72] | 17.7 | 29.3 | 35.4 | 35.2 | 32.8 | 21.3 | 30.7 | 34.0 | 34.1 | 32.5 |
| DeepCluster | 13.4 | 32.3 | 41.0 | 39.6 | 38.2 | 19.6 | 33.2 | 39.2 | 39.8 | 34.7 |

DeepCluster yields conv3–4 features that are comparable to those trained with ImageNet labels. This suggests that when the target task is sufficiently far from the domain covered by ImageNet, labels are less important.

4.4 Pascal VOC 2007

Finally, we do a quantitative evaluation of DeepCluster on image classification, object detection and semantic segmentation on PASCAL VOC. The relatively small size of the training sets on PASCAL VOC (2,500 images) makes this setup closer to a “real-world” application, where a model trained with heavy computational resources, is adapted to a task or a dataset with a small number of instances. Detection results are obtained using `fast-rcnn`²; segmentation results are obtained using the code of Shelhamer *et al.*³. For classification and detection, we report the performance on the test set of PASCAL VOC 2007 and choose our hyperparameters on the validation set. For semantic segmentation, following the related work, we report the performance on the validation set of PASCAL VOC 2012.

Table 2 summarized the comparisons of DeepCluster with other feature-learning approaches on the three tasks. As for the previous experiments, we outperform previous unsupervised methods on all three tasks, in every setting. The improvement with fine-tuning over the state of the art is the largest on semantic segmentation (7.5%). On detection, DeepCluster performs only slightly better than previously published methods. Interestingly, a fine-tuned random network performs comparatively to many unsupervised methods, but performs poorly if only FC6–8 are learned. For this reason, we also report detection and

² <https://github.com/rbgirshick/py-faster-rcnn>

³ <https://github.com/shelhamer/fcn.berkeleyvision.org>

Table 2: Comparison of the proposed approach to state-of-the-art unsupervised feature learning on classification, detection and segmentation on PASCAL VOC. * indicates the use of the data-dependent initialization of Krähenbühl *et al.* [31]. Numbers for other methods produced by us are marked with a †

| Method | Classification | | Detection | | Segmentation | |
|-----------------------------|----------------|-------------|-------------------|-------------|-------------------|-------------------|
| | FC6-8 | ALL | FC6-8 | ALL | FC6-8 | ALL |
| ImageNet labels | 78.9 | 79.9 | – | 56.8 | – | 48.0 |
| Random-rgb | 33.2 | 57.0 | 22.2 | 44.5 | 15.2 | 30.1 |
| Random-sobel | 29.0 | 61.9 | 18.9 | 47.9 | 13.0 | 32.0 |
| Pathak <i>et al.</i> [46] | 34.6 | 56.5 | – | 44.5 | – | 29.7 |
| Donahue <i>et al.</i> [15]* | 52.3 | 60.1 | – | 46.9 | – | 35.2 |
| Pathak <i>et al.</i> [45] | – | 61.0 | – | 52.2 | – | – |
| Owens <i>et al.</i> [44]* | 52.3 | 61.3 | – | – | – | – |
| Wang and Gupta [63]* | 55.6 | 63.1 | 32.8 [†] | 47.2 | 26.0 [†] | 35.4 [†] |
| Doersch <i>et al.</i> [13]* | 55.1 | 65.3 | – | 51.1 | – | – |
| Bojanowski and Joulin [5]* | 56.7 | 65.3 | 33.7 [†] | 49.4 | 26.7 [†] | 37.1 [†] |
| Zhang <i>et al.</i> [71]* | 61.5 | 65.9 | 43.4 [†] | 46.9 | 35.8 [†] | 35.6 |
| Zhang <i>et al.</i> [72]* | 63.0 | 67.1 | – | 46.7 | – | 36.0 |
| Noroozi and Favaro [42] | – | 67.6 | – | 53.2 | – | 37.6 |
| Noroozi <i>et al.</i> [43] | – | 67.7 | – | 51.4 | – | 36.6 |
| DeepCluster | 72.0 | 73.7 | 51.4 | 55.4 | 43.2 | 45.1 |

segmentation with FC6-8 for DeepCluster and a few baselines. These tasks are closer to a real application where fine-tuning is not possible. It is in this setting that the gap between our approach and the state of the art is the greater (up to 9% on classification).

5 Discussion

The current standard for the evaluation of an unsupervised method involves the use of an AlexNet architecture trained on ImageNet and tested on class-level tasks. To understand and measure the various biases introduced by this pipeline on DeepCluster, we consider a different training set, a different architecture and an instance-level recognition task.

5.1 ImageNet versus YFCC100M

ImageNet is a dataset designed for a fine-grained object classification challenge [51]. It is object oriented, manually annotated and organised into well balanced object categories. By design, DeepCluster favors balanced clusters and, as discussed above, our number of cluster k is somewhat comparable with the number of labels in ImageNet. This may have given an unfair advantage to

Table 3: Impact of the training set on the performance of DeepCluster measured on the PASCAL VOC transfer tasks as described in Sec. 4.4. We compare ImageNet with a subset of 1M images from YFCC100M [58]. Regardless of the training set, DeepCluster outperforms the best published numbers on most tasks. Numbers for other methods produced by us are marked with a †

| Method | Training set | Classification | | Detection | | Segmentation | |
|-----------------|--------------|----------------|------|-------------------|------|-------------------|------|
| | | FC6-8 | ALL | FC6-8 | ALL | FC6-8 | ALL |
| Best competitor | ImageNet | 63.0 | 67.7 | 43.4 [†] | 53.2 | 35.8 [†] | 37.7 |
| DeepCluster | ImageNet | 72.0 | 73.7 | 51.4 | 55.4 | 43.2 | 45.1 |
| DeepCluster | YFCC100M | 67.3 | 69.3 | 45.6 | 53.0 | 39.2 | 42.2 |

DeepCluster over other unsupervised approaches when trained on ImageNet. To measure the impact of this effect, we consider a subset of randomly-selected 1M images from the YFCC100M dataset [58] for the pre-training. Statistics on the hashtags used in YFCC100M suggests that the underlying “object classes” are severely unbalanced [28], leading to a data distribution less favorable to DeepCluster.

Table 3 shows the difference in performance on PASCAL VOC of DeepCluster pre-trained on YFCC100M compared to ImageNet. As noted by Doersch *et al.* [13], this dataset is not object oriented, hence the performance are expected to drop by a few percents. However, even when trained on uncured Flickr images, DeepCluster outperforms the current state of the art by a significant margin on most tasks (up to +4.3% on classification and +4.5% on semantic segmentation). We report the rest of the results in the supplementary material with similar conclusions. This experiment validates that DeepCluster is robust to a change of image distribution, leading to state-of-the-art general-purpose visual features even if this distribution is not favorable to its design.

5.2 AlexNet versus VGG

In the supervised setting, deeper architectures like VGG or ResNet [21] have a much higher accuracy on ImageNet than AlexNet. We should expect the same improvement if these architectures are used with an unsupervised approach. Table 4 compares a VGG-16 and an AlexNet trained with DeepCluster on ImageNet and tested on the PASCAL VOC 2007 object detection task with fine-tuning. We also report the numbers obtained with other unsupervised approaches [13, 64]. Regardless of the approach, a deeper architecture leads to a significant improvement in performance on the target task. Training the VGG-16 with DeepCluster gives a performance above the state of the art, bringing us to only 1.4 percents below the supervised topline. Note that the difference between unsupervised and supervised approaches remains in the same ballpark for both architectures (*i.e.* 1.4%). Finally, the gap with a random baseline grows for larger architectures,

Table 4: PASCAL VOC 2007 object detection with AlexNet and VGG-16. Numbers are taken from Wang *et al.* [64]

| Method | AlexNet | VGG-16 |
|----------------------------|-------------|-------------|
| ImageNet labels | 56.8 | 67.3 |
| Random | 47.8 | 39.7 |
| Doersch <i>et al.</i> [13] | 51.1 | 61.5 |
| Wang and Gupta [63] | 47.2 | 60.2 |
| Wang <i>et al.</i> [64] | – | 63.2 |
| DeepCluster | 55.4 | 65.9 |

Table 5: mAP on instance-level image retrieval on Oxford and Paris dataset with a VGG-16. We apply R-MAC with a resolution of 1024 pixels and 3 grid levels [59]

| Method | Oxford5K | Paris6K |
|----------------------------|-------------|-------------|
| ImageNet labels | 72.4 | 81.5 |
| Random | 6.9 | 22.0 |
| Doersch <i>et al.</i> [13] | 35.4 | 53.1 |
| Wang <i>et al.</i> [64] | 42.3 | 58.0 |
| DeepCluster | 61.0 | 72.0 |

justifying the relevance of unsupervised pre-training for complex architectures when little supervised data is available.

5.3 Evaluation on instance retrieval

The previous benchmarks measure the capability of an unsupervised network to capture class level information. They do not evaluate if it can differentiate images at the instance level. To that end, we propose image retrieval as a down-stream task. We follow the experimental protocol of Tolias *et al.* [59] on two datasets, *i.e.*, Oxford Buildings [48] and Paris [49]. Table 5 reports the performance of a VGG-16 trained with different approaches obtained with Sobel filtering, except for Doersch *et al.* [13] and Wang *et al.* [64]. This preprocessing improves by 5.5 points the mAP of a supervised VGG-16 on the Oxford dataset, but not on Paris. This may translate in a similar advantage for DeepCluster, but it does not account for the average differences of 19 points. Interestingly, random convnets perform particularly poorly on this task compared to pre-trained models. This suggests that image retrieval is a task where the pre-training is essential and studying it as a down-stream task could give further insights about the quality of the features produced by unsupervised approaches.

6 Conclusion

In this paper, we propose a scalable clustering approach for the unsupervised learning of convnets. It iterates between clustering with k -means the features produced by the convnet and updating its weights by predicting the cluster assignments as pseudo-labels in a discriminative loss. If trained on large dataset like ImageNet or YFCC100M, it achieves performance that are better than the previous state-of-the-art on every standard transfer task. Our approach makes little assumption about the inputs, and does not require much domain specific knowledge, making it a good candidate to learn deep representations specific to domains where annotations are scarce.

References

1. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: ICCV (2015)
2. Bach, F.R., Harchaoui, Z.: Diffrac: a discriminative and flexible framework for clustering. In: NIPS (2008)
3. Bautista, M.A., Sanakoyeu, A., Tikhoncheva, E., Ommer, B.: Cliqecnn: Deep unsupervised exemplar learning. In: Advances in Neural Information Processing Systems. pp. 3846–3854 (2016)
4. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: NIPS (2007)
5. Bojanowski, P., Joulin, A.: Unsupervised learning by predicting noise. ICML (2017)
6. Bojanowski, P., Joulin, A., Lopez-Paz, D., Szlam, A.: Optimizing the latent space of generative networks. arXiv preprint arXiv:1707.05776 (2017)
7. Bottou, L.: Stochastic gradient descent tricks. In: Neural networks: Tricks of the trade, pp. 421–436. Springer (2012)
8. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: CVPR (2016)
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915 (2016)
10. Coates, A., Ng, A.Y.: Learning feature representations with k-means. In: Neural networks: Tricks of the trade, pp. 561–580. Springer (2012)
11. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. vol. 1, pp. 1–2. Prague (2004)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
13. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
14. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning (2017)
15. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. arXiv preprint arXiv:1605.09782 (2016)
16. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: NIPS (2014)
17. Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., Courville, A.: Adversarially learned inference. arXiv preprint arXiv:1606.00704 (2016)
18. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. University of Montreal **1341**, 3 (2009)
19. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics New York (2001)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
21. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV (2015)
22. Huang, F.J., Boureau, Y.L., LeCun, Y., et al.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: CVPR (2007)
23. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)

24. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
25. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. arXiv preprint arXiv:1702.08734 (2017)
26. Joulin, A., Bach, F.: A convex relaxation for weakly supervised classifiers. arXiv preprint arXiv:1206.6413 (2012)
27. Joulin, A., Bach, F., Ponce, J.: Discriminative clustering for image co-segmentation. In: CVPR (2010)
28. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV (2016)
29. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
30. Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K., et al.: Crowdsourcing in computer vision. *Foundations and Trends® in Computer Graphics and Vision* **10**(3), 177–243 (2016)
31. Krähenbühl, P., Doersch, C., Donahue, J., Darrell, T.: Data-dependent initializations of convolutional neural networks. arXiv preprint arXiv:1511.06856 (2015)
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
33. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: ECCV (2016)
34. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998)
35. Liao, R., Schwing, A., Zemel, R., Urtasun, R.: Learning deep parsimonious representations. In: NIPS (2016)
36. Lin, F., Cohen, W.W.: Power iteration clustering. In: ICML (2010)
37. Linsker, R.: Towards an organizing principle for a layered perceptual network. In: NIPS (1988)
38. Mairal, J., Koniusz, P., Harchaoui, Z., Schmid, C.: Convolutional kernel networks. In: NIPS (2014)
39. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011)
40. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. *Artificial Neural Networks and Machine Learning* pp. 52–59 (2011)
41. Misra, I., Zitnick, C.L., Mitchell, M., Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: CVPR (2016)
42. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
43. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: ICCV (2017)
44. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV (2016)
45. Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B.: Learning features by watching objects move. *CVPR* (2017)
46. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
47. Paulin, M., Douze, M., Harchaoui, Z., Mairal, J., Perronin, F., Schmid, C.: Local convolutional features with unsupervised training for image retrieval. In: ICCV (2015)

48. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
49. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR (2008)
50. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015)
51. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *IJCV* **115**(3), 211–252 (2015)
52. de Sa, V.R.: Learning classification with unlabeled data. In: NIPS (1994)
53. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPR workshops (2014)
54. Shi, J., Malik, J.: Normalized cuts and image segmentation. *TPAMI* **22**(8), 888–905 (2000)
55. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
56. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *JMLR* **15**(1), 1929–1958 (2014)
57. Stock, P., Cisse, M.: Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. arXiv preprint arXiv:1711.11443 (2017)
58. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: The new data and new challenges in multimedia research. arXiv preprint arXiv:1503.01817 (2015)
59. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015)
60. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: CVPR (1991)
61. Van De Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. *TPAMI* **32**(9), 1582–1596 (2010)
62. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR* **11**(Dec), 3371–3408 (2010)
63. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV (2015)
64. Wang, X., He, K., Gupta, A.: Transitive invariance for self-supervised visual representation learning. arXiv preprint arXiv:1708.02901 (2017)
65. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: ICCV (2013)
66. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: ICML (2016)
67. Xu, L., Neufeld, J., Larson, B., Schuurmans, D.: Maximum margin clustering. In: NIPS (2005)
68. Yang, J., Parikh, D., Batra, D.: Joint unsupervised learning of deep representations and image clusters. In: CVPR (2016)
69. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
70. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)

71. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
72. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. arXiv preprint arXiv:1611.09842 (2016)
73. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS (2014)