Learning Discriminatory Deep Clustering Models

A. Alqahtani[1,2], X. Xie[1], J. Deng[1], and M. W. Jones[1]

[1] Department of Computer Science, Swansea University, Swansea, UK
http://csvision.swan.ac.uk
[2] Department of Computer Science, King Khalid University, Abha, Saudi Arabia

Abstract. Deep convolutional auto-encoder (DCAE) allows to obtain useful features via its internal layer and provide an abstracted latent representation, which has been exploited for clustering analysis. DCAE allows a deep clustering method to extract similar patterns in lower-dimensional representation and find idealistic representative centers for distributed data. In this paper, we present a deep clustering model carried out in presence of varying degrees of supervision. We propose a new version of DCAE to include a supervision component. It introduces a mechanism to inject various levels of supervision into the learning process. This mechanism helps to effectively reconcile extracted latent representations and provided supervising knowledge in order to produce the best discriminative attributes. The key idea of our approach is distinguishing the discriminatory power of numerous structures, through varying degrees of supervision, when searching for a compact structure to form robust clusters. We evaluate our model on MNIST, USPS, MNIST fashion, SVHN datasets and show clustering accuracy on different supervisory levels.

Keywords: Deep Convolutional Auto-Encoder · Embedded Clustering · Supervision.

1 Introduction

In recent years, deep learning methods have shown their robust ability in representation learning and achieved considerable success in many tasks. It transforms raw data into a more abstract representation. A Deep convolutional auto-encoder (DCAE) is a deep unsupervised model for representation learning. It maps inputs into a new latent space, allowing to obtain useful features via its encoding layer. This high-level representation provides beneficial properties that can support traditional clustering algorithms in demonstrating satisfying performance. DCAE has been exploited for clustering analysis, allowing such clustering algorithms to deal with an abstract latent representation in a low-dimensional space. Different approaches to unsupervised deep clustering have been developed utilizing deep neural networks. A detailed survey can be found in [10]. For instance,

DCAE with embedded clustering [2] is an unsupervised clustering method that simultaneously captures representative features and the relationships among images. In this procedure, the discriminative patterns are only discovered through certain parts or objects in an image in an unsupervised manner. The goal of this method is to learn feature representations and cluster assignments simultaneously, utilizing the strength of DCAE to learn high-level features. Two objective functions were utilized: one is embedded into a DCAE model to minimize the distance between features and their corresponding cluster centers, while the other one minimizes the reconstruction error of the DCAE. During optimization, all data representations are assigned to their new identical cluster centers and then cluster centers are updated iteratively allowing the model to achieve a stable clustering performance. The defined clustering objective, as well as the reconstruction objective, are simultaneously utilized to update parameters of transforming network.

Providing partial supervision to the clustering process, semi-supervised clustering aims to cluster a large amount of unlabeled data in the presence of a minimal supervision. Basu et al. [3] studied the effect of using a small amount of labeled data to generate initial seeds for K-means. Pedrycz et al. [11] also proposed a fuzzy clustering algorithm with partial supervision. Other works utilize pairwise constrained clustering method as semi-supervised process, which has been applied to partitioning clustering [14], hierarchical clustering [4], and density-based clustering [13]. Similarly, supervised clustering includes a supervisory scheme into the clustering process aiming to improve unsupervised clustering algorithms through exploiting supervised information [15]. Pedrycz et al. [12] presented fuzzy clustering algorithm with supervision that carried out in the presence of label information. Eick et al. [7, 6] introduced supervised clustering methods, which suppose that all obtained clusters hold ground truth labels aiming to identify class-uniform clusters. Al-Harbi et al. [1] also proposed a supervised clustering method by modifying the K-means algorithm to be used as a classifier.

Even though conventional semi-supervised and supervised clustering approaches have received a lot of attention, with the revolution of deep learning, limited attention has been paid to semi-supervised and supervised deep clustering models compared with unsupervised deep clustering. Therefore, providing a way to inject varying degrees of supervision into the body of the deep learning process and exploring the influence of adding supervision knowledge into a deep clustering model are worthwhile to understand discriminatory power obtained by patterns or provided by supervision components.

In this paper, we focus on a deep clustering model, where varying degrees of supervision can be injected into the body of the learning process. We propose a new version of DCAE to involve a supervision component. Injecting supervision allows us to experience different discriminatory powers, which can be provided by supervisory knowledge or obtained by data-driven discriminative attributes and examine the clustering performance through different levels of supervision. The proposed method is aimed at forming a kind of a structure that recon-

ciles structure discovered by the clustering process and structure provided by labeling patterns. This mechanism makes the features derived from the encoding layer are the best discriminative attributes. An available side of background knowledge along with representative patterns in latent space can be leveraged to find the best partitioning of data and maximize the purity of clusters. Experimental results illustrate the influence of adding supervision into the body of the learning process. In this study, we consider three different learning levels: supervised, semi-supervised and unsupervised. We evaluate our experimental models on MNIST, USPS, MNIST fashion, SVHN datasets and show clustering accuracy of our model through supervised, semi-supervised and unsupervised learning levels.

2 Method

The proposed approach is a DCAE with embedded clustering that is carried out in presence of varying degrees of supervision. It introduces a mechanism to inject various levels of supervision into the body of the learning process. This allows us to explore the leverage of supervised information into the performance of a deep clustering method. In this paper, we consider three different learning levels: supervised, semi-supervised and unsupervised. Each experimental model consists of combination objective functions. All objectives are simultaneously optimized.

2.1 DCAE with Embedded Clustering

DCAE is learned to capture representative features through its encoding layer by minimizing the reconstruction error using the Euclidean (L2) loss function.

$$E_1 = \frac{1}{2N} \sum_{i=1}^{n} \parallel x^i - y_i \parallel^2 \tag{1}$$

where y is a reconstructed image, and x is an original image.

Although DCAE learns an effective representation via its encoding layer, it does not explicitly force representation forming compact clustering. In [2], we proposed a DCAE with embedded clustering that learns feature representations and clusters assignments simultaneously. It embeds K-means clustering into a DCAE framework and minimizes the distance between data points and their assigned centers in the latent space as follows:

$$E_2 = \frac{1}{2N} \sum_{n=1}^{N} \parallel h^t(x_n) - c_n^* \parallel^2 \tag{2}$$

where N is the number of data examples, $h^t(*)$ denotes the encoded representation obtained at the t^{th} iteration, (x_n) is the n^{th} example in the dataset x. and c_n^* is the assigned centroids to the n^{th} example. For further detail of DCAE with embedded clustering, readers can refer to [2].

2.2 Architecture and Extended Output Layer

Using the extended version of the DCAE with embedded clustering method allows us to inject supervision and utilize its strength to obtain discriminative and robust features from the encoding layer and allows the deep clustering method to extract discriminative features and cluster assignments, simultaneously.

DCAE architecture consists of three convolutional layers. This is followed by two fully-connected layers, of which the second layer has 10 neurons. These are considered as hidden representations learned through the training process. A single fully-connected layer is followed by three deconvolutional layers as the decoding part. ReLU is utilized as the activation function. Table 1 has shown a detailed configuration of DCAE network architecture. Our extensions to this architecture are as follows. Firstly, instead of a reconstruction layer at the end of the DCAE, extra layers are added at the end of the network just after the reconstruction layer. This allows the passing of supervision knowledge across the learning process and also the examination of clustering performance with different discriminatory power that is provided by supervision or obtained from data-driven discriminative attributes. Secondly, the learned features given by the encoding layer are optimized to form compact and discriminative clusters using K-means, which minimizes the distance between a feature representation and their respective centroid. Thirdly, instead of only minimizing the reconstruction loss and cross-entropy loss, we iteratively optimize the mapping function of the encoding part and cluster centers to obtain more effective clustering.

Table 1: Detailed configuration of the DCAE network architecture used in the experiments.

Layer	***MNIST***	***USPS***	***MNIST Fashion***	***SVHN***
Convolutional	5 x 5 x 32	4x4x32	5x5x32	5x5x32
Convolutional	5x5x64	4x4x64	5x5x64	5x5x64
Convolutional	3x3x128	2x2x128	3x3x128	2x2x128
Fully Connected	1152	512	1152	2048
Fully Connected	10	10	10	10
Fully Connected	1152	512	1152	2048
Deconvolutional	3x3x128	2x2x128	3x3x128	2x2x128
Deconvolutional	5x5x64	3x3x64	5x5x64	5x5x64
Deconvolutional	5x5x32	3x3x32	5x5x32	5x5x32

In supervised and semi-supervised models, we have used same architectures, showing on Table. 1. Instead of a reconstruction layer at the end of the DCAE, we flatten the output of the reconstruction layer and feed them into a certain number of nodes in the last layer. The number of nodes depends on the task at hand, i.e. the number of provided classes (e.g. ten nodes for the supervised case and two nodes for the semi-supervised case). A softmax function is used for the

final prediction. The final architecture of our extended model for a DCAE with embedded clustering is presented in Fig.1.

Two forms of labels are used: **true labels** and **parent-class labels** to reflect two different levels of supervision. True labels are provided in supervised training process. Parent-class labels are used in semi-supervised deep clustering, that is true class labels are combined to form parent-class labels. For example, in clustering digit images using the proposed DCAE, the parent-class labels are defined as:

$$ParentLabel = \begin{cases} 0 & Labels < 5 \\ 1 & otherwise \end{cases} \tag{3}$$

The categorical cross-entropy function between network predictions and provided labels is defined as:

$$E_3 = -\sum_j t_{i,j} log(p_{i,j}) \tag{4}$$

where p is prediction, t is the provided label, i denotes the number of samples, and j denotes the class.

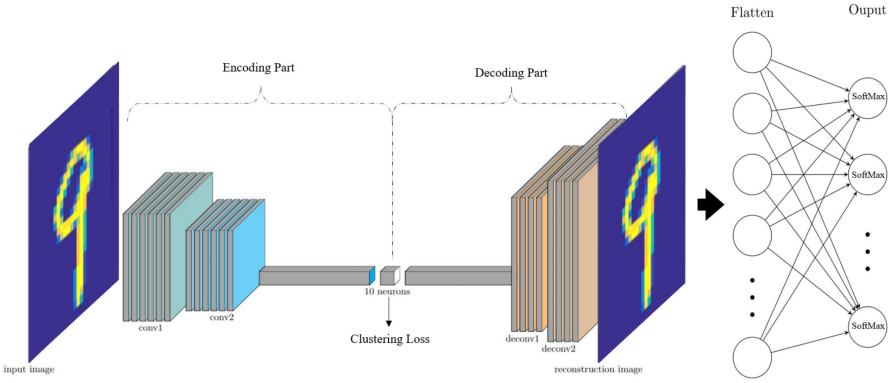

Fig. 1: The architecture of our proposed model.

In the DCAE hidden layer, encoded features are used to compute clustering loss function that minimizes the distance between data points and their corresponding cluster centers, Eqn. (2). The overall cost function is thus a combination of reconstructions loss $E1$, clustering residual in latent space $E2$, and categorical cross-entropy loss $E3$ that minimizes the classification error with either supervised or semi-supervised scheme:

$$\min_{W,b} E_1 + E_2 + E_3 \tag{5}$$

3 Experiments and Discussion

The proposed method was implemented using Keras and Theano in Python and evaluated on four different datasets including MNIST, USPS, MNIST fashion, and SVHN, which are the most commonly used datasets in the area of deep clustering. Specifications of these datasets are presented in Table. 2 . The model was trained end-to-end without involving any pre-training and fine-tuning procedures. All weights and cluster centers were initialized randomly. *Adam* optimizer was used where each batch contains 100 random shuffled images.

Table 2: Details of Datasets used in our experiments.

Dataset	***Examples***	***Classes***	***Image Size***	***Channels***
MNIST	70000	10	28x28	1
USPS	11000	10	16x16	1
MNIST Fashion	70000	10	28x28	1
SVHN	99289	10	32x32	3

For MNIST dataset, the experiments were performed using four different numbers of trained examples, i.e. 2000, 4000, 6000, 8000. We trained our supervised model using these settings with the same number of iteration. The comparative results are shown in Table. 3, which supports our hypothesis that a small amount of labeled data can add enough discriminative ability to unsupervised deep clustering. Note that the results are the accuracy of clustering not classification use reconstructed image.

Table 3: Number of trained samples and clustering accuracy.

Trained Examples	***Clustering Accuracy***
2000	94.24 %
4000	96.48 %
6000	97.52 %
8000	98.06 %

In order to visualize the impact of supervision in deep clustering, the t-SNE visualization method [9] was applied as a visual assessment to show adding supervision is able to guide the clustering task to obtain more appropriate data partitioning. Fig.2 shows the latent representation of our proposed methods in 2D space using different levels of supervisions, where color coding of the ground truth label are used to visualize the clustering results. This shows that adding a supervision component into DCAE with embedded clustering produces significantly more compact clusters. The learned features involves a supervised process have tighter structures and larger inter-cluster distances compared with semi-

supervised and unsupervised models. Injecting supervision into the learning process effectively reconciles data-driven-obtained representations and the provided supervisory knowledge to form the best partitioning of data and to maximize the purity of clusters. With the semi-supervised approach (Fig.2b), the clustering result show typical compact clusters, producing much better clustering results compared with unsupervised models (Fig.2c), which shows the learned features are sparse and not compacted. Fig.2d shows that the data distribution on latent space using normal DCAE which was trained only to optimize reconstruction error. Compared to Fig.2c which enforces compact representation on hidden layer, the clusters forming by normal DCAE have higher intra-cluster variance and lower inter-cluster difference. By adopting semi-supervised (see Fig.2b) and supervised (see Fig.2a), intra-cluster variances are reduced significantly while the inter-cluster distances are enlarged. Especially, less cluster outliers are observed in Fig.2a.

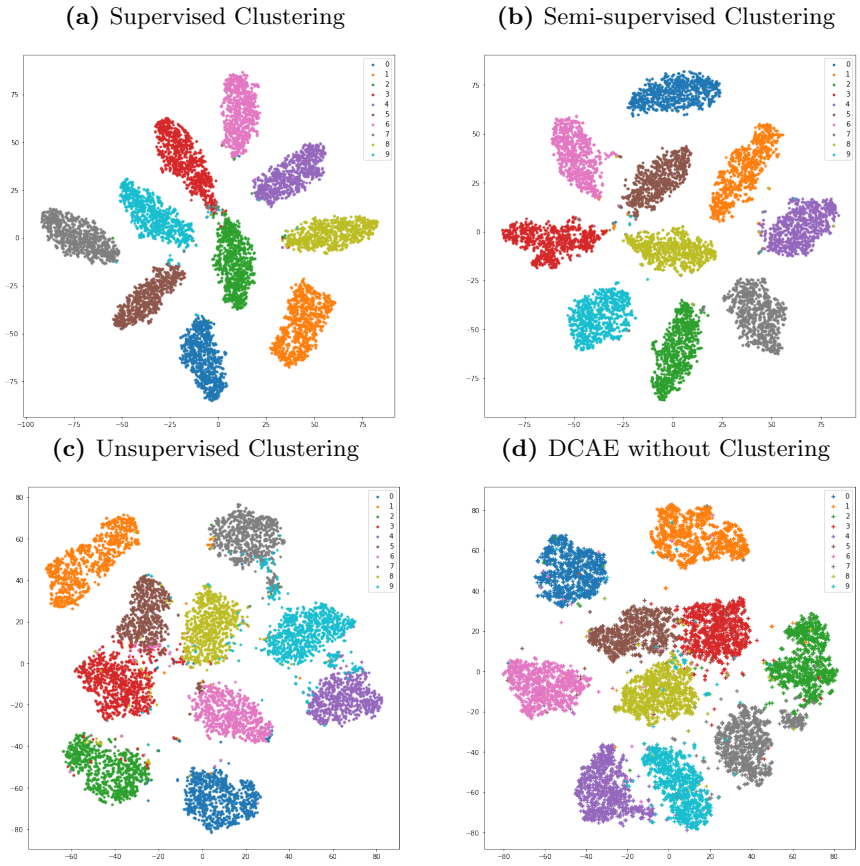

Fig. 2: Visualizations of latent representation for our method through a different supervisions levels on MNIST testing set.

In addition, we analyze the invariance properties of learned representation given different levels of supervision. We have trained five different models with varying degrees of supervision: supervised, semi-supervised with three different percentages of supervision (20%, 30%, 50%), and unsupervised. We apply a range of rotation-based transformations (rotate by 90^o, 180^o, 270^o, flip horizontally, flip horizontally and rotate by 90^o, 180^o, 270^o) to each image. We follow [5, 8] to measure the variance properties by calculating Mean Squared Error (MSE) between the features of the original images and the transformed ones. The result is shown in Fig.3. The figure compares the invariance properties of learned representation in five different models. Overall, the experiment empirically confirms that the features are more invariant when no supervision is provided. In other words, the features learned by the unsupervised model are more invariant compared to features learned with supervision.

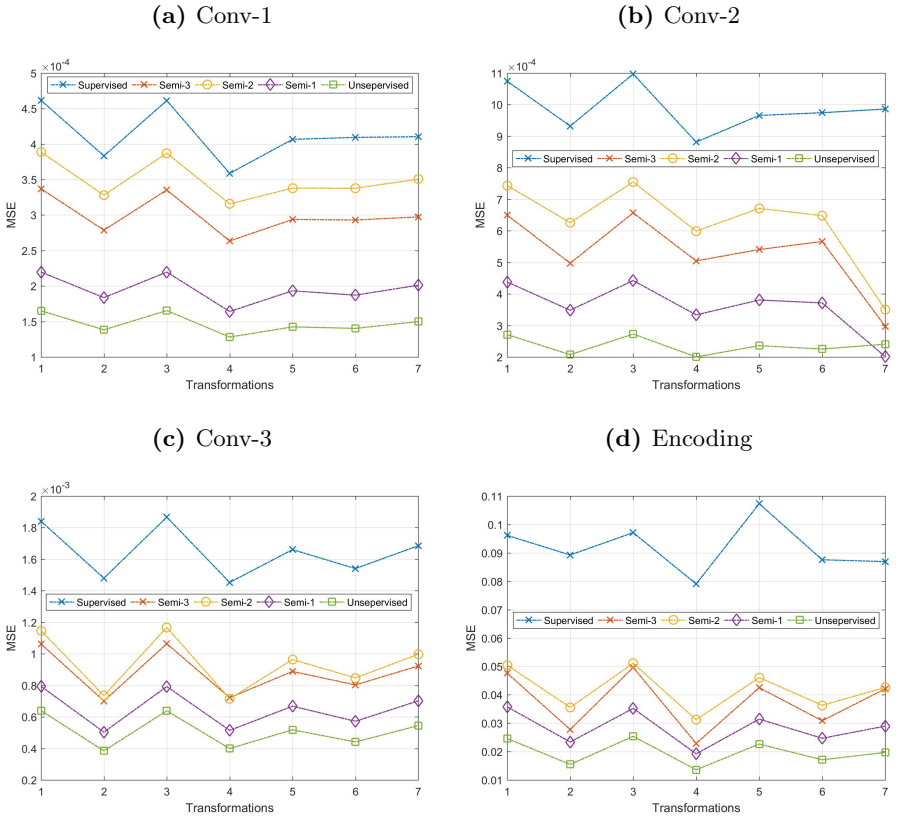

Fig. 3: Invariance properties of the learned representation in different layers from five different models.

We empirically evaluated the performance of representation learning of DCAE with different supervisory schemes by calculating the clustering accuracy against the true label. These experiments show a discriminative representation can form a kind of structure that reconciles structure discovered by clustering process and structure formed by labeling patterns. An available side of label information along with data-driven patterns were efficiently brought together to support the clustering process. Injecting true labels or partial supervision into the DCAE with an embedded clustering method allows the clustering algorithm to perform much better compared with the models that utilize an unsupervised learning process. Label consistency can add a discriminative power that clearly guides the clustering algorithm to obtain the best, most accurate compacted groups compared with data-driven discriminative attributes. Table 4 summarizes the results on four different datasets including MNIST, USPS, and more challenging ones, such as MNIST fashion and SVHN. Since the performances were evaluated on a classification task, the accuracy increasing with supervision knowledge enforced can be observed on both cases. Particularly for SVHN dataset, the accuracy is boosted more than two times when weak labels are provided. We argue that the common structures are not well formed without supervision where there are large variances in appearance and noisy in images are observed commonly in SVHN dataset.

Table 4: Comparison of clustering accuracy on four different datasets.

	MNIST	***USPS***	***MNIST fashion***	***SVHN***
Unsupervised [2]	92.14%	89.23%	60.42%	17.41%
Semi-supervised	97.77%	91.92%	63.59%	34.96%
Supervised	98.82%	95.06%	88.73%	92.40%

4 Conclusion

In the paper, we proposed a DCAE model which is capable of learning compact data representation that can be incorporated into different learning schemes, i.e. unsupervised, semi-supervised, supervised. We found that such supervision knowledge greatly helps to form discriminative transformations that are learned by the encoding part of a DCAE model and significantly improves the performance of clustering. It implies that the latent space has the potential to be used for image generation/ synthesis. The results also demonstrate that even weak or partial supervision knowledge can significantly improve the quality of deep clustering.

Acknowledgment

This work is supported by EPSRC EP/N028139/1.

References

1. Al-Harbi, S.H., Rayward-Smith, V.J.: Adapting k-means for supervised clustering. Applied Intelligence **24**(3), 219–226 (2006)
2. Alqahtani, A., Xie, X., Deng, J., Jones, M.: A deep convolutional auto-encoder with embedded clustering. In: IEEE ICIP. pp. 4058–4062 (2018)
3. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised clustering by seeding. In: ICML (2002)
4. Davidson, I., Ravi, S.: Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In: European Conference on Principles of Data Mining and Knowledge Discovery. pp. 59–70 (2005)
5. Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE T-PAMI **38**(9), 1734–1747 (2016)
6. Eick, C.F., Vaezian, B., Jiang, D., Wang, J.: Discovery of interesting regions in spatial data sets using supervised clustering. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) Knowledge Discovery in Databases: PKDD 2006. pp. 127–138. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
7. Eick, C.F., Zeidat, N., Zhao, Z.: Supervised clustering-algorithms and benefits. In: IEEE ICTAI. pp. 774–776 (2004)
8. Kavukcuoglu, K., Fergus, R., LeCun, Y., et al.: Learning invariant features through topographic filter maps. In: 2009 IEEE CVPR. pp. 1605–1612 (2009)
9. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)
10. Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., Long, J.: A survey of clustering with deep learning: From the perspective of network architecture. IEEE Access **6**, 39501–39514 (2018)
11. Pedrycz, W., Waletzky, J.: Fuzzy clustering with partial supervision. IEEE T-SMC-B **27**(5), 787–795 (1997)
12. Pedrycz, W., Vukovich, G.: Fuzzy clustering with supervision. Pattern Recognition **37**(7), 1339–1349 (2004)
13. Ruiz, C., Spiliopoulou, M., Menasalvas, E.: Density-based semi-supervised clustering. Data mining and knowledge discovery **21**(3), 345–370 (2010)
14. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al.: Constrained k-means clustering with background knowledge. In: ICML. vol. 1, pp. 577–584 (2001)
15. Zaghian, A., Noorbehbahani, F.: A novel supervised cluster adjustment method using a fast exact nearest neighbor search algorithm. Pattern Analysis and Applications **20**(3), 701–715 (2017)