Dimensionality Reduction Using Discriminative Autoencoders for Remote Sensing Image Retrieval

Mohbat, Tooba Mukhtar, Numan Khurshid, and Murtaza Taj

Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan
{16060073,19100210,15060051,murtaza.taj}@lums.edu.pk

Abstract. Advancements in deep learning techniques caused a paradigm shift in feature extraction for image perception from handcrafted methods to deep methods. However, these deep features if learned through unsupervised methods bear large memory footprints and are prone to the curse of dimensionality. Traditional feature reduction schemes involving aggregation of these learned visual descriptors may lead to loss of essential information necessary for their obvious discrimination. Therefore, this research studies various feature reduction techniques for remote sensing image features. We also propose an off-the-shelf deep discriminative network with dimensionality reduction (DAE-DR), exploiting stacked autoencoder based solution to abbreviate unsupervised features without significantly affecting their discriminative and regenerative characteristics. It is observed that the spatial dimensions encoded in the feature vector are more important than increasing the number of network filters for efficient image reconstruction. Validation of our approach has been tested for remote sensing image retrieval problem. Results demonstrate that our proposed network achieves 25 times reduction in feature size with only 0.8 times depletion of retrieval score.

Keywords: Remote sensing · Unsupervised features · Image retrieval · Deep Learning

1 Introduction

Developments in imaging technology have resulted in the extremely large datasets, however, learning any useful information from these datasets, particularly using modern deep learning architectures, require large amount of annotations. Although initiatives such as ImageNet challenge and those related to Autonomous Vehicles provide such annotated data, they are only limited to street level imagery. In many areas, such as remote sensing, there is a dearth of annotated datasets [6], Thus, there is a dire need of a method that allows unsupervised learning of features that are distinctive, posses reconstruction capability and are effectively compact.

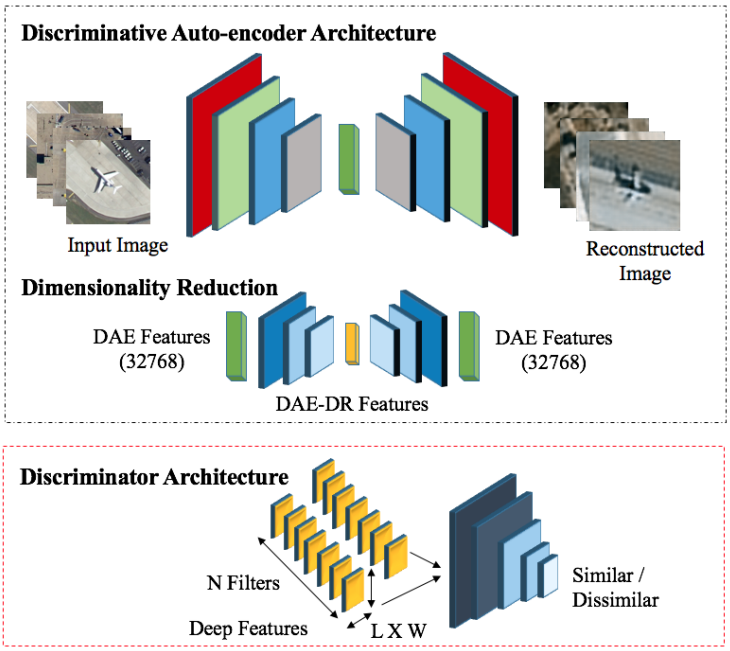


Fig. 1: Architecture of entire framework: The top box represents an Auto-encoder used to learn the features. The bottom box represents a network trained to discriminatively predict whether a pair of images arise from the same class or not based on the learned features.

To cultivate distinctiveness among unsupervised features, we adopted discriminative autoencoder network inspired from Generative Adverserial Networks (GANs) [4] and Siamese Networks [8] in previous work [11]. However, these learned features are usually high dimensional with large memory footprints which require huge storage capacity for big data applications, such as remote sensing image retrieval.

Dimensionality reduction could be considered as one of the possible solutions, employed through feature aggregation (by using global sum-pooling, max-pooling, and scaled sum-pooling) or selection of kernels from the activations of the learned network [5, 12]. However, firstly, these methods fail to perform on features learned through unsupervised learning methods. Secondly, these empirical techniques require unbounded set of experiments and even then does not guarantee compact feature representation.

In our previous work we proposed a Discriminative Autoencoder (DAE) architecture that takes as input high-dimensional features from the depth layer of autoencoder and project them onto a space that separates similar images from non-similar images (see Fig. 1)[11]. This work demonstrates a step-wise procedure to abbreviate the features acquired through deep autoencoder network

without significantly effecting their discriminative and regenerative characteristics.

Our approach leverages from the fact that autoencoders with linear activation are mathematically equivalent to Linear Principle Component Analysis (PCA) and those with non-linear activation (such as sigmoid) are equivalent to non-linear PCA. To prove the efficacy, we evaluated our approach on remote sensing image retrieval problem using benchmark datasets including University of California Merced Land Use/Land Cover (LandUse) [13] and High-resolution Satellite scene (SatScene) [3] containing 2100 and 1050 images, respectively..

2 Preliminaries

2.1 Discriminative Autoencoder (DAE)

In a dataset of images $X = \{x_1, x_2, \cdots\}$, our network maps the input image, x onto the feature space, f through several convolutional layers and activation functions $f = h_\theta(x) = r(Wx + b)$ and then using f reconstructs the output $x^{'}$ similar to the input image as $x^{'} = g_{\theta'}(f) = t(W^{'}f + b^{'})$. where h and g are encoder and decoder functions respectively while $\theta = \{W, b\}$ are encoder parameters, $\theta^{'} = \{W^{'}, b^{'}\}$ are decoder parameters for r,g being non-linear activation functions. By employing the mean squared error $L(x, x^{'}) = \|x - x^{'}\|^2$ as loss function, we optimize the parameters θ and $\theta^{'}$ as follows:

$$\theta^*, \theta^{'*} = \arg\min_{\theta,\theta'} \frac{1}{N} \sum_{i=1}^{N} L(x_i, x_i^{'}) \tag{1}$$

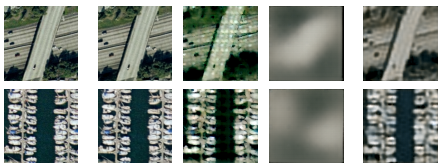
A pair of these image features (f_q, f_t) are then concatenated and given to the discriminator network $y^{'} = d((f_q, f_t), \theta_d)$ to compute the Bernoulli probabilities (match or unmatched), where d is a discriminator model and $y^{'}$ is classification probability. The parameters of d are optimized by using cross entropy loss function $L_d(y, y^{'}) = -\sum_{q,t} y log y^{'}$ as given in equation (2).

$$\theta_d^* = \arg\min_{\theta_d} \sum_{q,t} [L(y_i, d(h_\theta(x_q) * h_\theta(x_t))] \tag{2}$$

In our previous work [11], it has been demonstrated that the features f learned using residual autoencoder coupled with the discriminative metric learning scheme outperforms supervised features based approaches. However, these features suffers from curse of dimensionality.

2.2 Autoencoder vs PCA Relationship

We aim to obtain a transformation Φ that transform f to subspace $\tilde{f}$ as $\tilde{f} = \Phi_{\tilde{\theta}}(f, \tilde{W})$ and then from $\tilde{f}$ we aim to reconstruct the output $\tilde{x}^{'}$ as $\tilde{x}^{'} = g_{\tilde{\theta}}(\tilde{f}) = t(\tilde{W}\tilde{f} + \tilde{b})$ and compute similarity as $\tilde{y}^{'} = d(\{\tilde{f}_q, \tilde{f}_t\}, \tilde{\theta}_d)$. $\tilde{f}$ should be such that



(a) (b) (c) (d) (e)

Fig. 2: Visualization of reconstructed images from (a) Input (b) $8 \times 8 \times 512$ dimensional features of DAE (c) 1063 PCA basis (d) $1 \times 1 \times 1024$ dimensional encoder features (DAE-DR 1D) (e) $8 \times 8 \times 20$ dimensional encoder features (DAE-DR 2D).

it is a compact representation of f without any significant loss of information. In order to introduce energy conservation the transform should also be unitary i.e.

$$\|\tilde{f}\|^2 = \tilde{f}^H \tilde{f} = \Phi_{\tilde{\theta}}(f, \tilde{W})^H \Phi_{\tilde{\theta}}(f, \tilde{W}) = \|f\|^2 \tag{3}$$

where f^H is the hermitian conjugate of f. One such unitary transform is Eigen matrix of auto-correlation $R_{ff} = ff^H$ which form the basis of Principle Component Analysis (PCA). In order to learn the optimal feature vector, we exploited the relationship between PCA and auto-encoder basis. Mathematically, a linear autoencoder is defined as:

$$f_1 = W_1 \times X + b_1 \tag{4a}$$

$$\tilde{X} = W_2 \times f_1 + b_2 \tag{4b}$$

Where, W_1 and W_2 are weights, X is input and $\tilde{X}$ is reconstructed output. Minimizing the mean square cost function (equation 5) with respect to W_1, W_2, b_1, b_2, the problem reduces to optimization with respect to W_2 only, as given in equation (6).

$$min_{W_1,W_2,b_1,b_2} = \|X - (W_2(W_1X + b_1) + b_2)\|^2 \tag{5}$$

$$min_{W_2} = \|X^* - W_2 W_2^{\dagger} X^*\|^2 \tag{6}$$

Thus, by singular decomposition of W_2, it can be proven that the singular vectors of W_2 are actually the principle components of X. Consequently, PCA is equivalent to linear autoencoder whereas typical deep neural network based autoencoder with non-linear activation functions would be analogous to non-linear version of the PCA [2]. Therefore, the deep CNN autoencoder would learn feature space much better than PCA where PCA would help us to compute the optimal dimension of the space.

3 Methodology

3.1 Dimensionality Reduction in DAE via PCA

PCA helps to find the optimal dimension of space spanned by data but the challenge is the auto-correlation matrix which is computationally expensive. So,

instead of computing f as $f = \Phi^H x$ where $\Phi = \xi(R_{XX})$, and $\xi(.)$ returns the Eigen vectors, we compute Φ as $\Phi = F\xi(R_{F^H F})$ (using Sirovich and Kirby method [10]), i.e. by computing Eigen vectors of inner product of depth features instead of raw images, where $F = \{f_1, f_2, \cdots\}$. $\tilde{f}$ is then computed as:

$$\tilde{f} = \Phi^H f \tag{7}$$

Therefore, we compute the auto-correlation $R_{\tilde{F}^H \tilde{F}}$ of $\tilde{f}$ to identify the basis vectors that contain the maximum amount of energy. By reducing the feature dimension using PCA, from analysis of DAE features (32768 dimensions) on LandUse dataset, it has been found that 95% of the information lies in only 1063 principle components.

3.2 Dimensionality Reduction via DAE-DR Network

We modified our existing DAE architecture to DAE-DR network to learn the features with compact dimensions. The following three ways demonstrate the achieved modification for conversion of features from DAE to DAE-DR.

Pruning spatial dimensions of filters. By the introduction of 3 additional residual blocks in autoencoder, spatial dimension is reduced to 1×1 while increasing the number of filters to 1024, resulting in a 1D fine-grained feature vector (DAE-DR 1D). Nonetheless, as compared to PCA neither regeneration nor retrieval score were encouraging. It is quite obvious from Fig. 2(d) that reduction of spatial dimension of activation's results in loss of structural information and outputs a degraded reconstructed image, hence, confirming the idea presented in [11].

Pruning temporal dimensions of filters. Filters could be pruned by adapting "Try and learn" learning approach [7], converting DAE to DAE-DR. However, this method takes a lot of training time which exponentially increases with the complexity of network. Another way is to introduce a stack of layers which reduces the dimensions depth wise while keeping the spatial dimensions unchanged throughout. This technique ensures that the structural information is stored in the spatial dimension. However, the addition of depth in the network architecture produces a blurred regeneration.

Table 1: Regeneration loss: Averaged MSE on test set where training hyper-parameters were same for all models. PSNR averaged over 21 classes of LandUse.

Model/Scheme	Feature Size	MSE Loss	PSNR (dB)
DAE [11]	$(8, 8, 512)$	97.7	29.89
DAE (PCA)	1280	1114.34	6.150
DAE-DR 1D	$(1, 1, 1024)$	2179.32	16.541
DAE-DR 2D	$(8, 8, 20)$	636	21.192

Fig. 3: Qualitative Evaluation: Left most is query image and the remaining are top retrieved images.

Modification of existing DAE network. Additionally, another way is to modify the hidden layers of the original autoencoder network by manipulating the number of filters to produce the desired dimensional features. This approach yields the 2D compact features (DAE-DR 2D) with significant improvement in reconstruction as illustrated in Fig. 2(e).

The discriminator network for each of the three scenarios mentioned above has been modified in such a way that it accommodates the input feature dimension, preserving the overall architecture of the network.

4 Results and Discussion

4.1 Training and Evaluation

In order to evaluate the performance of reduced features with our previous results, all the training hyper-parameters were maintained as discussed in [11]. Data augmentation enabled the discriminator network to be robust to scaling, illumination and transnational invariances. For evaluation of all the approaches proposed in section 3, standard metrics discussed in [9] for remote sensing image matching were computed and a brief analysis has been provided in this section.

4.2 Analysis of Image Reconstruction

We trained three variants of auto-encoder networks and compared the regenerated images with [11]. Qualitative visual results demonstrated in Fig. 2 shows that the reconstruction of DAE-DR 2D features is smoother than reconstruction from PCA basis vectors. Moreover, the spatial compression of features results in the loss of structural information which degrades the reconstruction of the image. For quantitative evaluation, we compare the reconstruction MSE loss and Peak Signal to Noise Ration (PSNR). From Table 1, it can also be noticed that the MSE loss of DAE-DR 1D feature is almost 20 times higher than the loss of DAE. It can also be clearly analyzed that with the decrease in the feature dimension, the quality of the reconstructed images is impaired. Hence, for an effective reconstruction of the images local spatial information is extremely crucial.

Table 2: Comparative evaluation of our proposed approach for feature dimension reduction where it should be noted that despite having smaller feature size, our approach outperform hand-crafted features and is comparable with supervised deep features.

Feature Type	Features	Feature Size	ANMRR↓
LandUse Dataset			
Hand-crafted	LBP RGB [9]	54	0.751
	SIFT (VLAD) [14]	25600	0.649
	SIFT (FV) [9]	40960	0.639
Deep-supervised	NetVLAD [1]	4096	0.406
	SatResNet-50 [9]	2048	0.239
Deep-Unsupervised	DAE	32768	0.090
	DAE (PCA)	1063	0.591
	DAE-DR 1D	**1024**	0.495
	DAE-DR 2D	1280	**0.417**
SatScene Dataset			
Hand-crafted	LBP RGB [9]	54	0.664
	SIFT (VLAD) [14]	25600	0.649
	SIFT (FV) [9]	40960	0.552
Deep-supervised	NetVLAD [1]	4060	0.371
	SatResNet-50 [9]	2048	0.207
Deep-Unsupervised	DAE	32768	0.060
	DAE (PCA)	**804**	**0.473**
	DAE-DR FG 1D	1024	0.495
	DAE-DR 2D	1280	0.50

4.3 Analysis of Remote Sensing Image Matching

In order to evaluate the performance of metrics used for remote sensing image matching, we computed the values of Average Normalized Modified Retrieval Rank (ANMRR) and Mean Average Precision (mAP) [9]. Subjective evaluation by observing Fig. 3 clearly shows that the top 10 retrieved images mostly belong to the same class, however, the retrieved images are sometimes confused with visually similar images of different classes e.g. forest with rivers and over-head with highway class. The previously proposed unsupervised features outperforms supervised features in terms of lower ANMRR and higher mAP values which is evident from Table. 2 and a comparative analysis of features represented in Fig. 4 and Fig. 5. In our case, even with 25 times reduction in feature size, the performance is still comparable to hand-crafted approaches and competing with other supervised approaches e.g. NetVLAD [1]. As described in Table 2, the ANMRR value of DAE is comparatively better as compare to other DAE-DR unsupervised feature approaches for LandUse dataset. Even this approach outperforms other hand-crafted approaches in terms of ANMRR and feature size.

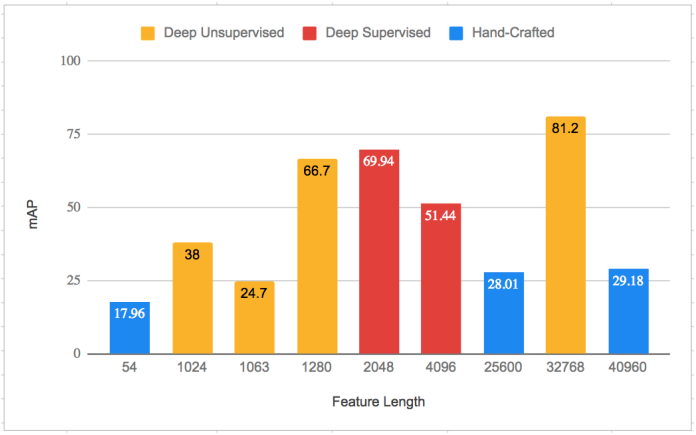


Fig. 4: Comparison between different feature sizes and their mAP scores for **LandUse** dataset.

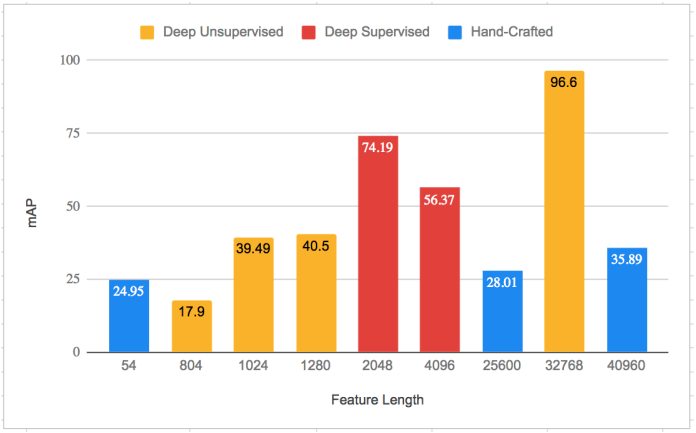


Fig. 5: Comparison between different feature sizes and their mAP scores for RS **SatScene** dataset.

Such significant differences in metric values demonstrates the effectiveness and superiority of our proposed feature size for the problem of remote sensing image retrieval using unsupervised features. By exploiting the local spatial and global semantic information, the proposed feature length outperforms the baseline sizes.

5 Conclusion

This paper introduces a novel unsupervised dimensionality reduction network after thoroughly studying some of the systematic methods of reducing unsupervised feature dimension including PCA. Through experiments we have shown

that our proposed network DAE-DR 2D is able to achieve comparable content based image retrieval results from a significantly smaller feature vector. While a larger number of feature maps are required to obtain accurate retrieval results, we show that by retraining the spatial information and discarding the redundant filters it is possible to produce an optimal size image descriptor employing discriminative autoencoder.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
2. Bourlard, H., Kamp, Y.: Auto-association by multilayer perceptrons and singular value decomposition. Biological cybernetics **59**(4-5), 291–294 (1988)
3. Dai, D., Yang, W.: Satellite image classification via two-layer sparse coding with biased image representation. IEEE Geoscience and Remote Sensing Letters **8**(1), 173–176 (2011)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27. pp. 2672–2680. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
5. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision. pp. 241–257. Springer (2016)
6. Haklay, M., Weber, P.: Openstreetmap: User-generated street maps. Ieee Pervas Comput **7**(4), 12–18 (2008)
7. Huang, Q., Zhou, K., You, S., Neumann, U.: Learning to prune filters in convolutional neural networks. arXiv preprint arXiv:1801.07365 (2018)
8. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: Deep Learning Workshop, International Conference on Machine Learning. Lille, France (2015)
9. Napoletano, P.: Visual descriptors for content-based retrieval of remote-sensing images. International Journal of Remote Sensing **39**(5), 1–34 (2018). https://doi.org/10.1080/01431161.2017.1399472
10. Sirovich, L., Kirby, M.: Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America **4**(3), 519–524 (1987)
11. Tharani, M., Khurshid, N., Taj, M.: Unsupervised deep features for remote sensing image matching via discriminator network. arXiv preprint arXiv:1810.06470 (2018)
12. Xia, G.S., Tong, X.Y., Hu, F., Zhong, Y., Datcu, M., Zhang, L.: Exploiting deep features for remote sensing image retrieval: A systematic investigation. arXiv preprint arXiv:1707.07321 (2017)
13. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. pp. 270–279. ACM (2010)
14. Zhou, W., Newsam, S., Li, C., Shao, Z.: Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. Remote Sensing **9**(5), 489 (2017)