DIMENSIONALITY REDUCTION IN AUTOENCODERS FOR UNSUPERVISED LEARNING

Tooba Mukhtar and Mohbat Tharani
Advisor: Murtaza Taj

Computer Vision Lab
Department of Computer Science,
LUMS School of Science and Engineering

ABSTRACT

With recent advancements in deep learning, paradigm of feature extraction for image perception has been shifted from handcrafted methods to deep learning algorithms. The set of discriminative features are usually prone to the curse of dimensionality if learned through unsupervised methods. Current feature reduction schemes aggregate learned visual descriptors which may lead to loss of essential information necessary for differentiating the features. This work demonstrates a systematic procedure to abbreviate the features acquired through deep autoencoder network without significantly effecting their discriminative and regenerative characteristics. We also show that spatial dimension encoded in feature vector is crucial for image reconstruction and cannot be replaced by increasing the number of distinctive activations learned through convolution layers. Validation of the approach has been tested for remote sensing image retrieval problem. Results demonstrate that our approach achieves $25\times$ reduction in feature size with only $0.8\times$ reduction in retrieval score. This paper also presents a detailed analysis to prune deep and/or wide convolutional neural network models by eliminating redundant features (or filters). Previous studies have shown that over-sized deep neural network models tend to produce a lot of redundant features that are either shifted version of one another or are very similar and show little or no variations; thus resulting in filtering redundancy.

***Index Terms*—** 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, however they are only limited to street level imagery. In many areas, such as remote sensing, there is a dearth of annotated datasets [1], 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.

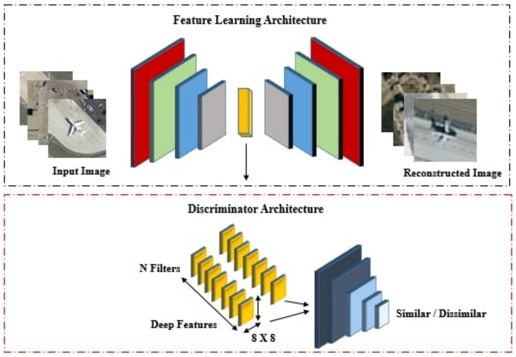

Fig. 1: Architecture of entire framework: The top box represents an Auto-encoder used to learn the features. The bottom box represents a network trained to discriminatively predict whether a pair of images arise from the same class or not based on the learned features.

To cultivate distinctiveness among unsupervised features, we adopted discrimination network in previous work [2] inspired from Generative Adverserial Networks (GANs) [3] and Siamese Networks [4]. However, these learned features are usually high dimensional with large memory footprints and they may work fine for smaller datasets but they require huge storage spaces for big data applications, such as remote sensing image retrieval.

Dimensionality reduction could be considered as one of the possible solution, employed through aggregation of features (by using global sum-pooling, max-pooling, and scaled sum-pooling), selection of kernels from the activations of the learned network [5, 6] or by pruning redundant activation's from the network. Firstly, these methods perform well for supervised learning, however, they fail to produce reconstructed images for unsupervised autoencoders. Secondly, these empirical techniques require unbounded set of experiments and does not guarantee compact feature representation.

In our previous work we proposed a Discriminative Autoencoder (DAE) architecture which includes a discriminator network 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). 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 University of California Merced Land Use/Land Cover (LandUse) [7] and High-resolution Satellite scene (SceneSet) [8].

2. METHODOLOGY

2.1. Discriminative Autoencoder (DAE)

Consider a dataset of images $X = \{x_1, x_2, \cdots\}$. Our **discriminative autoencoder** network takes an image x as input and maps it onto a feature space representation 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 $x^{'} = g_{\theta'}(f) = t(W^{'}f + b^{'})$. h and g are encoder and decoder functions respectively. $\theta = \{W, b\}$ are encoder parameters, $\theta^{'} = \{W^{'}, b^{'}\}$ are decoder parameters and r,g are non-linear activation functions. By employing the traditional loss function, mean squared error $L(x, x^{'}) = \|x - x^{'}\|^2$, we optimize the parameters θ and $\theta^{'}$ as follows:

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

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).

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

In our previous work [2], it has been demonstrated that the features f learned using residual auto-encoder coupled with the discriminative distance learning scheme to differentiate between similar and dissimilar images outperforms supervised schemes. However, these features suffers from curse of dimensionality.

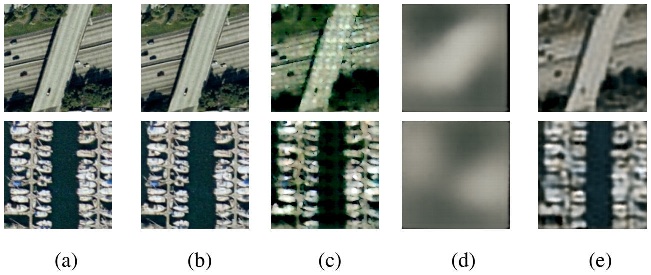
(a) (b) (c) (d) (e)

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

2.2. Dimensionality Reduction via PCA

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 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.

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

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).

Space spanned by PCA components could also be learned using linear autoencoder where the resultant vectors may not be orthogonal as in case of PCA basis [9]. In order to learn the optimal feature vector, we exploited the relationship between PCA and auto-encoder basis. Mathematically, linear autoencoder is defined as:

$$f_1 = W_1 \times X + b1 \qquad (4a)$$

$$\tilde{X} = W_2 \times f_1 + b_2 \qquad (4b)$$

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).

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

$$min_{W_2} = \|X^* - W_2 W_2^\dagger X^*\|^2 \qquad (6)$$

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 [10]. 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.

Since PCA requires computation of auto-correlation matrix R_{XX} which is computationally expensive so, instead of computing f as $f = \Phi^H x$ where $\Phi = \xi(R_{XX})$, and function $\xi(.)$ which returns the Eigen vectors, we compute Φ as $\Phi = F\xi(R_{F^H F})$ (using Sirovich and Kirby method [11]), 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. From PCA analysis of DAE features (32768 dimensions) on LandUse dataset, it has been found that 95% of the information lies in only 1063 principle components. PCA analysis is only performed for LandUse and SatScene datasets containing 2100 and 1050 images, respectively, using the network pre-trained on 400K images of GTCrossView dataset.

.

2.3. Dimensionality Reduction via Autoencoder

We redesigned our DAE architectures of autoencoder to learn desired dimensional features. In the following three ways, the modification for conversion of features from course grained (CG) to fine grained (FG) has been achieved.

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 feature vector. Nonetheless, as compared to PCA neither regeneration nor retrieval score were encouraging. It is quite obvious from Fig. 2 that reduction of spatial dimension of activations results in loss of structural information and outputs a degraded reconstructed image, hence, confirming the idea presented in [2].

Table 1: Regeneration loss: Averaged mean square error on test set where training hyper-parameters were same for all models. DAE CG: Discriminative Auto-Encoder Course Grained, DAE FG: Discriminative Auto-Encoder Fine Grained

Model/Scheme	Feature Size	MSE Loss
DAE CG [2]	$(8, 8, 512)$	97.7
DAE CG (PCA)	1063	1114.34
DAE FG 1D	$(1, 1, 1024)$	2179.32
DAE FG 2D	$(8, 8, 20)$	636

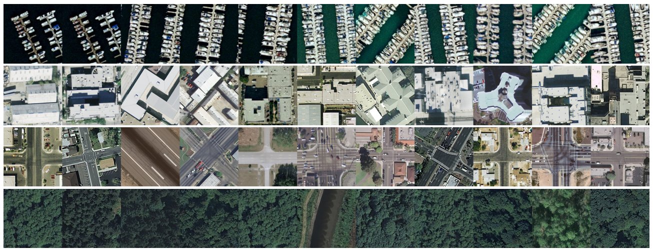

Fig. 3: Qualitative Evaluation: Left most is query image and the remaining are top retrieved images.

Pruning temporal dimensions of filters. One of the ways to prune the number of filters is to use "Try and learn" learning algorithm for filter pruning [12], converting 512 coarse-grained filters to 20 fine-grained filters. However, this method takes a lot of training time which exponentially increases with the complexity of network. Moreover, we also tried to use sub-layers in the network in order to reduce the number of filters through learning but its reconstruction score was not satisfactory as well.

Modification of existing autoencoder network. Another way is to modify the middle layer of the original autoencoder network to produce the desired number of filters. This approach yields the 2D fine-grained features of dimension $(8 \times 8 \times 20)$ with significant improvement in reconstruction as illustrated in Fig. 2.

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

3. DUPLICATE FILTERS IN NEURAL NETWORKS

Studies have shown that deep neural network models tend to produce a lot of redundant features that are extremely similar to each other; thus resulting in filtering redundancy. This phenomenon is prevalent in networks and increases with the number of layers. We observed the emergence of duplicate filters, study the factors that affect their concentration and compare existing network reducing operations. In the analysis, we focus on two ways in which filters in neural networks may be redundant (1) if they have negligibly small values (i.e. the filters have weight vectors with low l1-norm), or (2) if their functionality is mimicked by another filter. The latter may be quantified as high cosine similarity between the weight vectors of two filters.

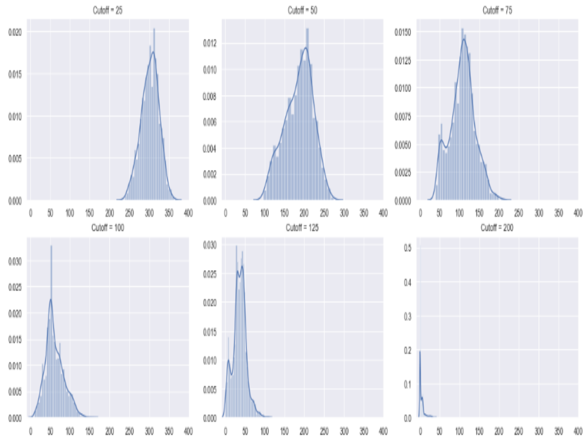

Fig. 4: Distribution of remaining activations of 8 x 8 x 512 for l1 based thresholds.

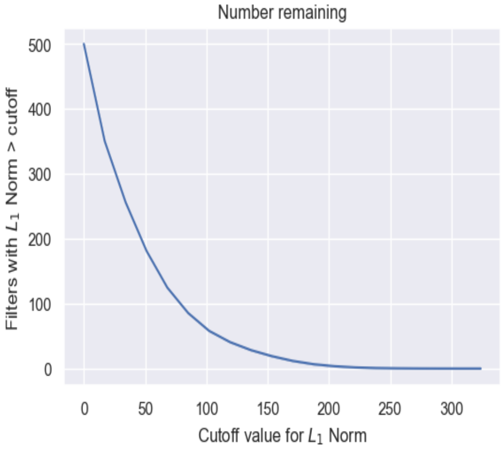

Fig. 5: Number of remaining filters with the increase in cutoff threshold for activation pruning.

3.1. Distribution of redundant filters

L1-norm based aggregation: To eliminate filters based on the low-l1 norm, we selected a random of 500 norm's of image activation's(8 x 8 x 512) to plot their distributions. The resulting histograms of the norms showed the average number of image activation's that were covered under several norm thresholds. This strategy gave an idea about the average value of norm across all the activation's. The aim was to find the optimal cutoff point for all the images such that if the input is N x 8 x 8 x 512 then the output will be N x 8 x 8 x P where P is the number of activation's which fulfil the criteria: **norm** $>$ **cutoff** and N is the number of images. After finding the optimal cutoff value of norm, each image had a different number of activation's pass the aforementioned criteria. Hence we looked at the distributions of the remaining activation's for all images. Figure.4 illustrates this idea. The horizontal axis shows the count of the number of remaining activation's after pruning. For instance, when the cutoff is 25 the range of remaining activation's after being pruned is around 250 to 350. This is the range across all the image activation's. Figure.5 shows that as we increase the norm threshold, the number of remaining filters shifts leftwards (towards 0). This makes sense since a higher norm cutoff means that less activation's pass the criteria and the distribution moves left and also shrinks in variance. The smallest feature size for 8 x 8 x 512 that we could get was 8 x 8 x 50. These plots give a visual indication of low-norm redundancy. This approach reiterates that dimensionality reduction can be achieved through filter pruning/aggregation. The experiments were carried on each subsequent layer of the encoder network for the LandUse dataset.

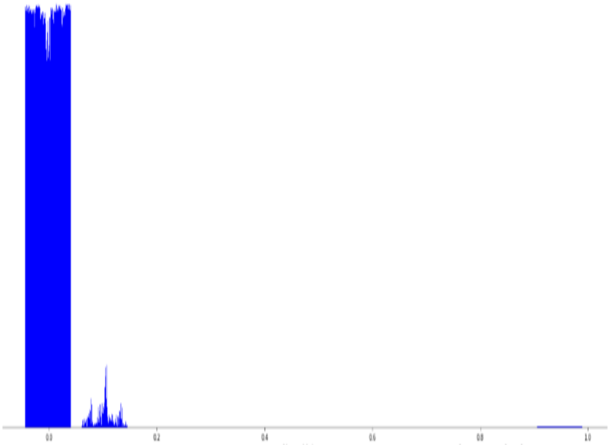

Fig. 6: Frequency distribution for cosine similarity of filters (512 x 3 x 3 x 512).

Cosine Similarity based aggregation: Having filters in a network with high cosine similarity means that the vectors are both pointing in the same direction, with a possible difference in scale. This reflects the similarity in the filters proving the existence of duplicate activation's which can be traced back to duplicate filters. The experiments were performed on the filters of each layer of the encoder network for the LandUse dataset. Fig 6. shows the distribution of cosine similarity between filters of the last layer of encoder. Our analysis shows that majority of the values lie close to 0 which means that most of the filters learned by the auto encoder network are not redundant as is only an insignificant amount of filters with values close to 1.0.

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 hyperparameters were maintained as discussed in [2]. Data augmentation enabled the discriminator network to be robust to scaling, illumina-

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 outperforms hand-crafted features and is comparable with supervised deep features.

Feature Type	Features	Feature Size	ANMRR	mAP	P@5	P@10	P@50	P@100
LandUse Dataset								
Hand-crafted	LBP RGB [13]	54	0.751	18.0	58.7	49.8	28.1	19.6
	Dense SIFT (VLAD) [14]	25600	0.649	28.0	74.9	65.3	38.2	28.1
	Dense SIFT (FV) [13]	40960	0.639	29.2	75.3	66.3	39.1	28.5
Deep-Supervised	NetVLAD [15]	4096	0.406	51.4	83.0	78.6	61.6	49.0
	SatResNet-50 [13]	2048	0.239	69.9	92.1	89.0	77.2	64.4
Deep-Unsupervised	DAE CG	32768	0.090	81.2	100	99.2	99.2	87.4
	DAE CG (PCA)	1063	0.591	24.7	59.6	54.9	45.5	39.9
	DAE FG 1D	**1024**	0.495	38.0	63.1	59.0	54.0	49.6
	DAE FG 2D	1280	0.417	43.9	66.7	66.5	63.6	57.2
SatScene Dataset								
Hand-crafted	LBP RGB [13]	54	0.664	25.0	50.3	44.0	26.3	19.4
	Dense SIFT (VLAD) [14]	25600	0.649	28.0	74.9	65.3	4.20	28.1
	Dense SIFT (FV) [13]	40960	0.552	35.9	71.3	62.8	36.2	25.0
Deep-Supervised	NetVLAD [15]	4060	0.371	56.4	82.5	78.4	64.4	52.2
	SatResNet-50 [13]	2048	0.207	74.2	92.1	90.6	80.9	68.0
Deep-Unsupervised	DAE CG	32768	0.060	96.6	100	100	94.3	52.0
	DAE CG (PCA)	**804**	0.473	17.9	65.0	62.0	52.1	37.2
	DAE FG 1D	1024	0.495	17.1	54.2	54.4	51.2	30.4
	DAE FG 2D	1280	0.50	17.6	60.8	60.0	50.7	29.9

tion and translational invariances. For evaluation of all the approaches proposed in section 2, standard metrics discussed in [13] for remote sensing image matching were computed and a brief analysis has been provided in this section.

4.2. Analysis of Image Reconstruction

Auto-encoder network has been trained to project the images onto a feature space as well as projecting them back onto the image space. These images mapped onto a feature space are reconstructed and an averaged mean square loss is computed between the original image and its reconstruction. Some qualitative visual results for these images have been demonstrated in Fig. 2 which shows that the reconstruction of $8 \times 8 \times 20$ is smoother than reconstruction from PCA basis vectors. According to Table 1, the reconstruction MSE loss of $1 \times 1 \times 1024$ dimensional feature is almost 20 times higher than the loss of discriminative autoencoder coarse-grained features (DAE CG). The spatial compression of features results in the loss of structural information which degrades the reconstruction of the image. It could be clearly analyzed that as the dimension of the features decreases, the quality of the reconstructed images is impaired. Hence, for an effective reconstruction of the images local spatial information is extremely crucial.

4.3. Dimensionality Reduction with unsupervised learning

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), Mean Average Precision (mAP), and Precision@K measures [13]. 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 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. 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 [15]. As described in Table 2, the ANMRR value of two dimensional discriminative autoencoder fine-grained (DAE CG 2D) is comparatively better as compare to other fine-grained unsupervised feature approaches for LandUse dataset. Even this approach outperforms other hand-crafted approaches in terms of mAP and feature size. Such significant differences in metric values demonstrates the effectiveness and superiority of our proposed feature size for the problem of unsupervised remote sensing image retrieval. By exploiting the local spatial and global semantic information, the proposed feature length outperforms the baseline sizes.

4.4. Dimensionality Reduction with filter aggregation

We have shown that duplication of filters occurs in CNNs as increasing the number of filters at a layer results in more duplicates. In the pruning stage, some of the weak filters can be discarded along with the corresponding feature maps. We focused on two questions: 1). How to prune redundant CNN filters 2). How to prune filters to control the trade off between network performance and filter aggregation.

5. CONCLUSION

This paper introduces a systematic method of reducing unsupervised feature dimension by exploiting the relationship between auto-encoder and PCA. Through experiments we have shown that it is possible to achieve comparable content based image retrieval results from a significantly smaller feature vector with respect to unsupervised models. 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. The auto-encoder framework was specifically modified to transform the image data into a compressed feature space. A substantial amount of experimental results on publically available remote sensing datasets validate the supremacy of our proposed approach. Moreover, this paper introduced an analysis approach to demonstrate the fact that there exists unnecessary CNN filters which if removed can reduce the feature size significantly thereby reducing the number of network parameters as well.

6. REFERENCES

[1] Mordechai Haklay and Patrick Weber, "Openstreetmap: User-generated street maps," *Ieee Pervas Comput*, vol. 7, no. 4, pp. 12–18, 2008.

[2] Mohbat Tharani, Numan Khurshid, and Murtaza Taj, "Unsupervised deep features for remote sensing image matching via discriminator network," *arXiv preprint arXiv:1810.06470*, 2018.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. 2014, pp. 2672–2680, Curran Associates, Inc.

[4] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Deep Learning Workshop, International Conference on Machine Learning*, Lille, France, 2015.

[5] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus, "Deep image retrieval: Learning global representations for image search," in *European Conference on Computer Vision*. Springer, 2016, pp. 241–257.

[6] Gui-Song Xia, Xin-Yi Tong, Fan Hu, Yanfei Zhong, Mihai Datcu, and Liangpei Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *arXiv preprint arXiv:1707.07321*, 2017.

[7] Yi Yang and Shawn Newsam, "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*. ACM, 2010, pp. 270–279.

[8] Dengxin Dai and Wen Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 173–176, 2011.

[9] Elad Plaut, "From principal subspaces to principal components with linear autoencoders," *arXiv preprint arXiv:1804.10253*, 2018.

[10] Hervé Bourlard and Yves Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological cybernetics*, vol. 59, no. 4-5, pp. 291–294, 1988.

[11] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America*, vol. 4, no. 3, pp. 519–524, 1987.

[12] Qiangui Huang, Kevin Zhou, Suya You, and Ulrich Neumann, "Learning to prune filters in convolutional neural networks," *arXiv preprint arXiv:1801.07365*, 2018.

[13] Paolo Napoletano, "Visual descriptors for content-based retrieval of remote-sensing images," *International Journal of Remote Sensing*, vol. 39, no. 5, pp. 1–34, 2018.

[14] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sensing*, vol. 9, no. 5, pp. 489, 2017.

[15] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.