AN EXPLAINABLE CONVOLUTIONAL AUTOENCODER MODEL FOR UNSUPERVISED CHANGE DETECTION

Luca Bergamasco[1,2], Sudipan Saha[1,2], Francesca Bovolo[1]*, Lorenzo Bruzzone[2]

[1] Fondazione Bruno Kessler, Trento, Italy - (lbergamasco, saha, bovolo)@fbk.eu
[2] University of Trento, Trento, Italy - (luca.bergamasco, sudipan.saha, lorenzo.bruzzone)@unitn.it

KEY WORDS: Multi-temporal Analysis, Change Detection, Deep Learning, Transfer Learning, Autoencoder, Explainable Artificial Intelligence

ABSTRACT:

Transfer learning methods reuse a deep learning model developed for a task on another task. Such methods have been remarkably successful in a wide range of image processing applications. Following the trend, few transfer learning based methods have been proposed for unsupervised multi-temporal image analysis and change detection (CD). Inspite of their success, the transfer learning based CD methods suffer from limited explainability. In this paper, we propose an explainable convolutional autoencoder model for CD. The model is trained in: 1) an unsupervised way using, as the bi-temporal images, patches extracted from the same geographic location; 2) a greedy fashion, one encoder and decoder layer pair at a time. A number of features relevant for CD is chosen from the encoder layer. To build an explainable model, only selected features from the encoder layer is retained and the rest is discarded. Following this, another encoder and decoder layer pair is added to the model in similar fashion until convergence. We further visualize the features to better interpret the learned features. We validated the proposed method on a Landsat-8 dataset obtained in Spain. Using a set of experiments, we demonstrate the explainability and effectiveness of the proposed model.

1. INTRODUCTION

Multi-temporal image analysis is one of the most popular research topics in remote sensing. It is important for monitoring phenomena like natural disasters (Adams, 2004) and urbanization (Del Frate et al., 2008). In the last 15 years, many new satellite based sensors have been launched by space agencies, thus increasing the number of available sensors periodically orbiting the Earth. This has improved the availability of multi-temporal data with higher revisit period. Currently, images are available from different imaging modalities (passive/active) and different spectral, spatial, and temporal resolutions (Bovolo, Bruzzone, 2015). This has resulted in strong increase in development of novel multi-temporal image analysis methods, especially aiming towards unsupervised Change Detection (CD) (Celik, 2009) (Bovolo, Bruzzone, 2015). However, the unsupervised CD methods in the literature often need to be largely modified to account for differences in acquisition sensor and resolution (Bovolo, Bruzzone, 2015).

A possible solution to design CD frameworks that would not need large modification for different sensors/resolutions, comes in the form of deep learning. Deep learning, a highly data-driven paradigm, has obtained state-of-the-art performance in almost all computer vision tasks (LeCun et al., 2015). Being data-driven, often deep learning based frameworks can be suitably used in new tasks by merely changing the training data. Its superior performance can be attributed to its excellent capability to extract semantically rich visual features (Zhou et al., 2014) and robust feature representation. Owing to its success, they have been rapidly adopted by the remote sensing community too (Zhang et al., 2016) (Ball et al., 2017). Deep learning based methods have been proposed for many remote sensing tasks, including semantic segmentation of aerial images (Maggiori et al., 2016) (Volpi, Tuia, 2017) and hyperspectral image analysis (Ma et al., 2015).

While data driven and supervised, deep learning based methods have been adopted for unsupervised applications via transfer learning (Ozbulak et al., 2016)(Penatti et al., 2015)(Huang et al., 2017). Exploiting transfer learning, Saha *et. al.* (Saha et al., 2019a) proposed Deep Change Vector Analysis (DCVA) for change detection in Very High spatial Resolution (VHR) optical images. DCVA uses a pre-trained network as bi-temporal deep feature extractor. DCVA has been extended for other imaging modalities with few modifications, e.g., High spatial Resolution (HR) images (Saha et al., 2019c), multi-sensor images (Saha et al., 2019b). Despite success of transfer learning based DCVA framework in unsupervised CD, it suffers from two limitations:

1. The pre-trained model needs to be trained for a supervised classification, i.e., training of the pre-trained model requires labeled single-time patches.
2. The pre-trained model needs to be trained on images collected from a geographical location similar in behaviour to the bi-temporal images for change detection.
3. How deep features trained for a classification task behave in the context of CD on bi-temporal images, is not completely interpretable/explainable.

To alleviate the first and second limitations, Bergamasco *et. al.* (Bergamasco et al., 2019) proposed a variant of DCVA in which a deep convolutional autoencoder (CAE) (Guo et al., 2017) is trained in unsupervised way on the unlabeled patches extracted from the same geographic location as the bi-temporal changes. It is noteworthy that unlabeled data are available in abundance for all geographical location for the sensors like Landsat-8 and Sentinel-2, that come with free data access policy. Subsequently, the CAE model is used as a bi-temporal deep feature extractor in a DCVA framework (Saha et al., 2019a). However, this

*Corresponding author.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1513-2020 | © Authors 2020. CC BY 4.0 License.

model still suffers from the aforementioned third limitation of a typical DCVA model. Though the CAE model is trained on same geographic location as bi-temporal images, its transferability from an image reconstruction task to CD is not completely interpretable/explainable. CAE is trained for reconstructing the training patches, thus not for CD. It is important to understand the relationship between the features learned by the CAE and their role in CD.

In last couple of years, we have seen efforts in the deep learning community to better explain/interpret the deep learning models, that were considered black-box algorithms few years back. Some of the works towards investigating the explainability of deep learning models is found in (Xue, Chuah, 2019)(Roscher et al., 2019)(Angelov, Soares, 2019). Motivated by this, in this paper we aim to design an explainable CAE model that better explains its usefulness for CD. To make an explainable model, we take inspiration from the greedy layerwise training proposed in the seminal work of Bengio *et. al.* (Bengio et al., 2007). We train a pair of encoder-decoder layers for image reconstruction. We evaluate the fitness of learned features for CD by ranking them according to their variance on the bi-temporal scene (Saha et al., 2019a). We retain only those features that are useful for CD. We add more layers in iterative fashion and after adding each layer, the fitness of the features for CD is evaluated. Thus features retained in each layer of the CAE are only those that explains their usefulness for the CD task. This sets it apart from the paradigm of DCVA where features are learned on a completely different task and reused in CD. While the proposed model benefits from transfer learning by learning a model for scene reconstruction, the features learned simultaneously accounts for their use in CD. We validated our method on a Landsat-8 dataset that shows a burned area in Spain. The model can be adapted for other imaging modalities/resolutions by changing training dataset for CAE.

This paper is organized into following sections. The proposed CAE based CD framework is presented in section 2. Experimental results are presented in section 3. We conclude the paper and discuss possible future works in section 4.

2. PROPOSED METHOD

Let X_1, X_2 be two images taken over the same region at time t_1, t_2, respectively. Let the set of all pixels in the bi-temporal scene be represented by the set of classes $\Omega = \{\Omega_c, \omega_{nc}\}$. The proposed explainable CAE method aims to distinguish the changed pixels Ω_c from the unchanged ones ω_{nc}. Let us assume that a dataset of unlabeled patches $\mathbf{X} = \{\mathbf{x_i}, \forall \mathbf{i} = \mathbf{1}, ..., \mathbf{I}\}$ is also available from the same geographical region as X_1 and X_2. This dataset is used to train a CAE architecture with $2L$ layers, consisting of L encoder layers and same number of decoder layers. The training process is achieved one encoder-decoder layer pair at a time. After training each pair, features explainable for CD are selected by extracting the features from the encoder layer of the model representing the bi-temporal scene. Only these features are retained and the training process is continued by adding one encoder-decoder layer pair at a time in a greedy manner (Bengio et al., 2007) until $2L$ layers are trained. After training, X_1 and X_2 are compared in a DCVA fashion to distinguish Ω_c from ω_{nc}. The proposed CD framework is shown in Figure 1.

In section 2.1, we briefly review basic CAE model. We detail the training of a CAE in section 2.2. In section 2.3, we describe the process of using the trained CAE for CD on the bi-temporal scene.

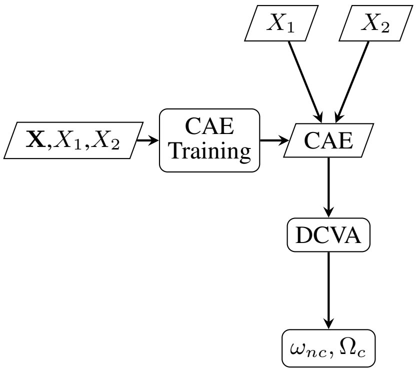

Figure 1. Proposed CD framework

2.1 Basic Convolutional-Autoencoder model

Autoencoders learn useful features from data in unsupervised way by learning to encode the data and further decode them back to the original input. However, autoencoder cannot capture the spatial context that is essential for the image analysis. On the contrary, standard deep architectures exploiting convolutional layers, such as Convolutional Neural Networks (CNNs), are capable of capturing the spatial context. However, CNNs cannot be trained in unsupervised manner. Convolutional Autoencoders (CAEs) (Guo et al., 2017) merge the capability of autoencoders to automatically learn features by input data, with the capability of convolutional layers to study the context of images and extract spatial context features from input images.

The CAE encoder down-samples the spatial resolution of input images and increases the number of features extracted by each layer. On the contrary, the CAE decoder up-samples the spatial resolution and reduces the number of features. The decoder block is usually composed of deconvolutional layers, which up-sample the spatial resolution by convolving the data with their kernels. The output of both encoder and decoder layers $l \in 2L$ of an input sample $\mathbf{x_i} \in \mathbf{X}$ is given by $\mathbf{h_{i,l}} = \sigma(\mathbf{W_l h_{i,l-1}} + \mathbf{b_l})$, where $\mathbf{W_l}$ and $\mathbf{b_l}$ are the weights and the biases of layer l, and σ represents the ReLU activation function. The input to the first layer is defined by $\mathbf{h_{i,0}} = \mathbf{x_i}$.

2.2 Training CAE

The CAE model is initiated with 2 layers (1 encoder and 1 decoder). The training is performed in two steps: 1) training of single encoder-decoder layer pair (section 2.2.1) ; 2) selecting explainable features from the trained encoder layer (section 2.2.2). After initialization of the first encoder-decoder pair, new encoder-decoder layer pairs are added as described in section 2.2.3 until the number of layers reaches $2L$. Finally, the trained model is fine-tuned as described in section 2.2.4. The CAE training process is shown in Figure 2.

2.2.1 Training single layer A single encoder-decoder layer pair of CAE essentially consists of a convolutional layer and a deconvolutional layer. The CAE is trained for image reconstruction on the dataset $\mathbf{X}$. For patches $\mathbf{x_i} \in \mathbf{X}$, the CAE reconstructs $\mathbf{x'_i} \in \mathbf{X'}$. They are compared using a sum squared error (SSE).

$$SSE = \sum_{i=1}^{\mathbf{I}} (\mathbf{x'_i} - \mathbf{x_i})^2 \tag{1}$$

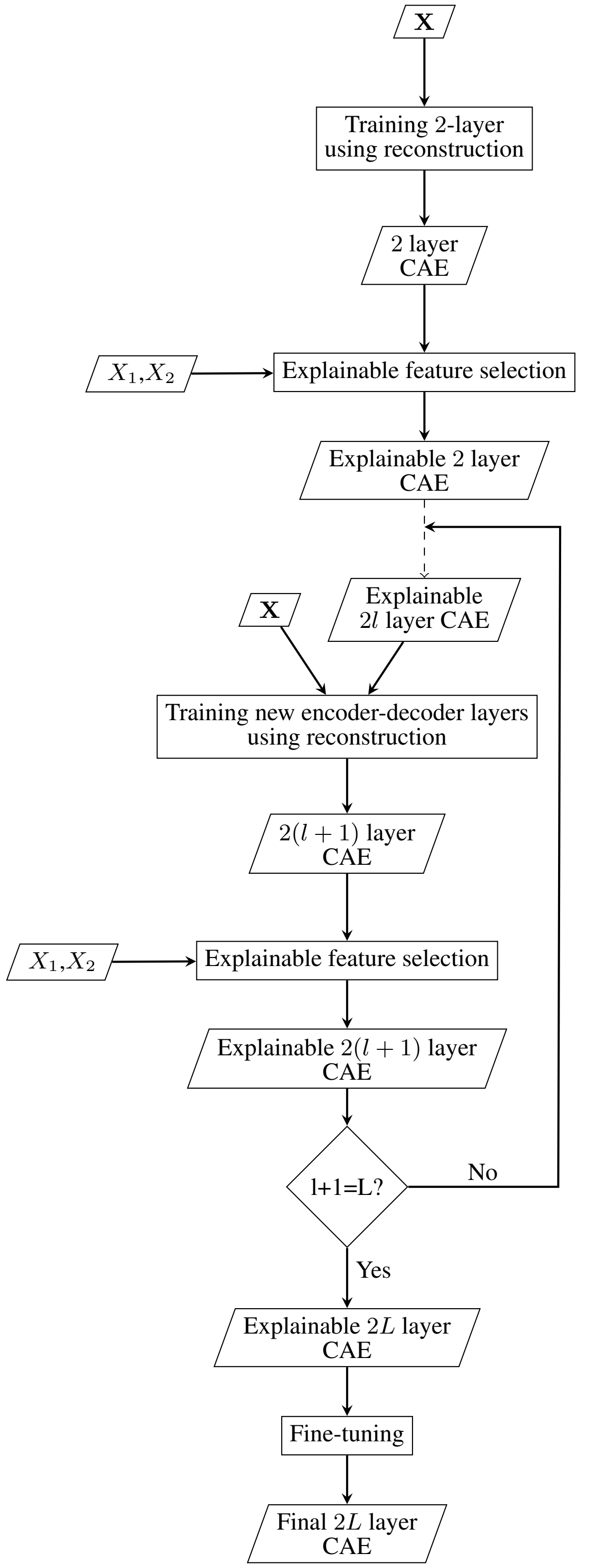
X
Training 2-layer using reconstruction
2 layer CAE
X_1,X_2
Explainable feature selection
Explainable 2 layer CAE
X
Explainable $2l$ layer CAE
Training new encoder-decoder layers using reconstruction
$2(l+1)$ layer CAE
X_1,X_2
Explainable feature selection
Explainable $2(l+1)$ layer CAE
l+1=L?
No
Yes
Explainable $2L$ layer CAE
Fine-tuning
Final $2L$ layer CAE

Figure 2. Explainable training of the CAE - training of 1st encoder and decoder layers involves only $\mathbf{X}$ and bi-temporal images X_1,X_2. Training of subsequent layers also involves the already trained network. A $2l$ layer CAE consists of l encoder layers and l decoder layers

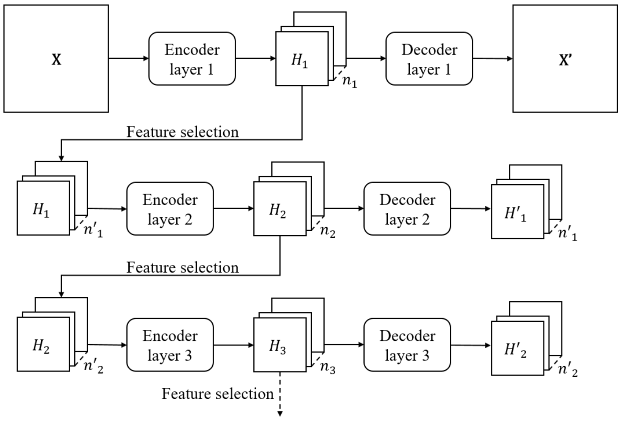

Figure 3. The dynamic Greedy layer-wise training process. Training of single encoder-decoder layer pair, feature selection and augmenting new encoder-decoder layer pair.

The CAE is trained for E epochs and the training process is performed by the back-propagation process.

2.2.2 Choosing explainable features Once an encoder, decoder layer pair of the CAE is trained for patch reconstruction on $\mathbf{X}$, the trained CAE is applied on the bi-temporal images X_1 and X_2 and feature-wise differences are taken. This feature-wise differences are computed by retrieving the features H_1 and H_2 from the first hidden layer of the CAE. H_1 and H_2 can be defined as

$$H_i = \sigma(W^1 * X_i + b^1) \tag{2}$$

where $i = 1, 2$, W^1 and b^1 are the weights and the biases of the first hidden layer of the CAE, respectively, and σ is the activation function. The bi-temporal features H_1 and H_2 are compared by using a squared error $D = (H_2 - H_1)^2$ to highlight the difference between the bi-temporal images X_1 and X_2, coded in the first hidden layer. Inspired by (Saha et al., 2019a), we assume that features capturing relevant change information tend to result in a D with higher standard deviation than those less responsive to change. In a bi-temporal scene, change has a small probability to occur. Thus, even in the case that a change occurs, it only partially affects the considered scene. In this context, and after computing D, features not affected by change show values that all tend to zero. These features have very low standard deviation values, and thus they are considered as non informative. Features affected by change shows both values that tend to zero for the portion being not affected by the change, and values far from zero for the portion being affected. Accordingly, the standard deviation of the features in the latter case tends to be greater than the one of the former ones, and therefore we consider them as informative. To optimally detect the changes in a scene, we have to select only the most informative features. Thus in a greedy fashion, for an encoder layer composed of n_l features, we select $n'_l = p \cdot n_l$ features, where p is the fraction of features to keep, with the highest standard deviation of D. These n'_l features are considered the most informative features of the layer l. Only the informative features are retained in the CAE and the training process is continued for successive layers.

The feature selection process is not solely based on image reconstruction task, which is completely different from CD task. Rather, the proposed method includes the bi-temporal images in the training of the model. In addition to training for image reconstruction, features from each layer are selected based on how well they describe the difference in bi-temporal images X_1 and X_2. The features chosen in this fashion are explainable for their application in change detection. This is in stark contrast to DCVA (Saha et al., 2019a)(Saha et al., 2019c). Even though, they used a feature selection during the CD process, their model itself was not trained for CD. Hence the use of those models for CD is not completely explainable. On the other hand, only those features are retained in the proposed explainable CAE model that can be interpreted to be useful for CD.

2.2.3 Augmenting new layers To add new encoder-decoder layer pair to the CAE, the weights of the already trained layers are frozen. After adding the new layer pair, the CAE is trained for image reconstruction on the dataset $\mathbf{X}$, in similar fashion as described in section 2.2.1. The augmented layers aim to obtain an output as similar as possible to the input. After training the CAE for E epochs, a feature selection process is applied again on this newly added layer, as outlined in section 2.2.2. This process is iterated until all $2L$ layers of the stacked CAE model are trained. The value of L is a trade-off between the size of the receptive field of the features (LeCun et al., 2015) and the computation cost. As L is increased, the CAE learns features that look at larger spatial region. However, computational cost increases with L as well. The maximum possible value of L is also restricted by the size of the patches in $\mathbf{X}$. The greedy layerwise training process is demonstrated in Figure 3.

2.2.4 Fine-tuning CAE After training all $2L$ layers, The pre-trained model is fine-tuned for E_{ft} epochs. Unlike the previous steps, the fine-tuning step is performed on all the layers simultaneously. This step helps to further train the whole model and improve its overall performance.

2.3 Using CAE for CD

Once the training process is over, we exploit the features learned by the CAE as a bi-temporal deep feature extractor in a DCVA framework (Saha et al., 2019a) for CD on the bi-temporal images X_1 and X_2. X_1 and X_2 are separately processed through the trained CAE which is used as a deep feature extractor. Deep features are extracted from a set of layers of the trained CAE network to form a deep feature hypervector. The set of layers are chosen from the encoder as decoder mainly learns to reconstruct the input scene. We obtain layerwise difference of the deep feature vectors and they are upsampled to the same spatial size of the input images. Following this, a DCVA analysis exploting Otsu's thresholding (Otsu, 1979) is performed to retrieve change maps for each layer. The multi-resolution change maps are then processed by the detail-preserving multi-scale approach proposed by Bovolo (Bovolo, Bruzzone, 2005) to distinguish Ω_c from ω_{nc}.

3. EXPERIMENTAL RESULTS

3.1 Dataset

The test dataset is acquired over an area near Granada, Spain from the Landsat-8 sensor on June 30^{th}, 2015 (Figure 4(a)) and July 16^{th}, 2015 (Figure 4(b)). They show an area of pixels 720 × 810 pixels. The area is impacted by fire between the two acquisitions. Reference CD map is shown in Figure 4(c).

3.2 Results

For training the CAE, we used patches of size 64 × 64 pixels and six of the eight spectral bands of Landsat-8 images. Based on some preliminary experiments, we set $L = 3$. We trained the model from scratch by using, for the encoder, a starting number of filters equal to $n_l = 32, 64, 128$. The filters of decoder mirror the encoder ones. By following the greedy layer-wise method, we trained each layer l for $E = 10$ epochs and fine-tune the model by using $E_{ft} = 100$. E and E_{ft} were set on the basis of some preliminary experiments. Figure 4(d) shows the CD map obtained by the proposed method for $p = 0.3$, i.e., when 30% features are retained in each layer. Table 1 shows the performance of the proposed method for different values of the parameter p. To test the effectiveness of the feature selection (FS), experiments were conducted by reducing p from 1.0 to 0.3,0.25,0.2, where the case of $p = 1.0$ considers both semantic features related to the changes and non informative ones. As one can see, considering non informative features decreases the probability to detect the changes. In the case of $p = 1.0$, the Missed Alarm (MA) rate is higher than in the cases where $p = 0.3, 0.25, 0.2$. The MA rate sharply decreases when p goes from 1.0 to 0.3. However, the MA rate slightly decreases when p is decreased to 0.25 from 0.3. MA rate increases sharply as p is decreased to 0.2 from 0.25. This proves that we need to do a trade-off during the FS. If p is too low, some informative features risk to be not considered leading a decrease of the capability of the model to detect the changes. False alarm (FA) rate produced by the method is negligible, as shown in Table 1. We compare our method with the Symmetric Convolutional Coupling Network (SCCN) (Liu et al., 2016), the Change Vector Analysis (CVA)(Bruzzone, Prieto, 2000) and the Stacked-Autoencoder (SAE)(Xu et al., 2013) based method. We provided the quantitative results in Table 1. Most deep-learning-based CD methods in the literature are supervised. Hence, a comparison of the proposed method with them is unfair. DCVA (Saha et al., 2019a) is unsupervised, however, designed for VHR images, and therefore a direct comparison with DCVA is not possible. Moreover, DCVA requires a pre-trained network, and, to the best of our knowledge, there is no such pre-trained network publicly available for the low-resolution multi-spectral images. Our method outperforms SCCN, CVA and SAE based method as shown in Table 1.

Method	OA	FA rate	MA rate
CVA	94.01%	4.45%	21.61%
Autoencoder (Xu et al., 2013)	76.45%	21.61%	42.06%
SCCN (Liu et al., 2016)	88.72%	10.97%	14.28%
Proposed ($p = 1.0$) no FS	95.16%	2.47%	27.45%
Proposed ($p = 0.2$)	95.16%	2.76%	24.74%
Proposed ($p = 0.25$)	95.36%	3.30%	17.45%
Proposed ($p = 0.30$)	96.64%	1.83%	17.96%

Table 1. The Overall Accuracy (OA), the Missed Alarm (MA) rate, and the False Alarm (FA) rate of the state-of-the-art methods and the proposed method.

3.3 Feature visualization

To further demonstrate the explainability of the proposed method, we visualize the difference image generated by three features that are retained by the proposed method in Figure 5 (a)-(c). It is clear that those features have learned useful semantic features related to the changes occurred in the considered scene. Whereas, the feature shown in Figure 5 (d) did not and is among the ones discarded by the method. A feature-by-feature analysis pointed out that the model correctly excludes the features similar to the one in Figure5(d) while training the CAE and keeps the ones that are useful for CD.

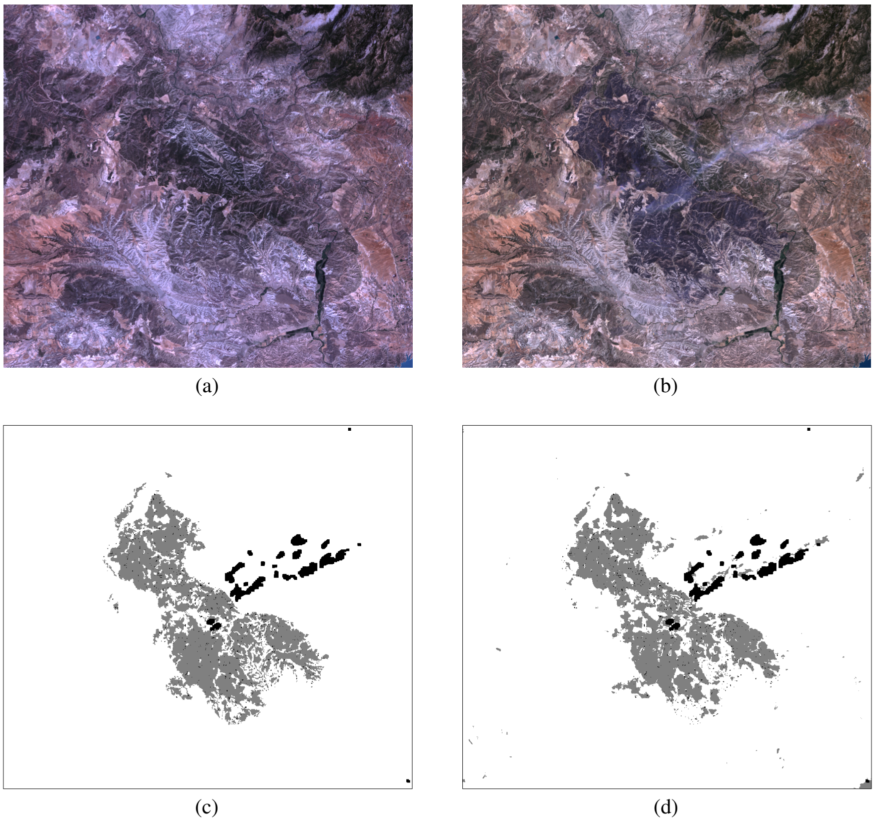

Figure 4. Spain Landsat-8 dataset: (a) pre-change image (RGB); (b) post-change image (RGB); (c) reference map; (d) CD map obtained by the proposed method. Unchanged pixels appear in white, changed pixels in grey. No reference data could be retrieved for pixels in black.

4. CONCLUSION

In this paper, we proposed an explainable CAE model for unsupervised change detection. While the proposed method takes advantage of transfer learning by learning to reconstruct a dataset of patches, the learned features are further selected based on a standard-deviation criterion after each layer is trained. Thus, the features retained in the CAE are explainable in terms of usefulness in CD. This is further confirmed by the visualization of the features. We tested the method on a Landsat-8 burned area dataset that confirms the effectiveness of the proposed method. The method can be extended for other imaging modalities/resolutions with simple modifications. This work is a step towards designing a more explainable deep learning model for CD. In future, we will perform extensive performance analysis by varying value of L. Additionally, we would like to extend CAE to distinguish between different kinds of change and understand how the explanation of different kinds of change can be incorporated CAE training process.

ACKNOWLEDGEMENTS

The authors are grateful to Miguel Angel Belenguer-Plomer (Department of Geology, Geography, and Environment of Universidad de Alcalá) for providing the reference map used for the experiments.

REFERENCES

Adams, B., 2004. Improved disaster management through post-earthquake building damage assessment using multitemporal satellite imagery. *Proceedings of the ISPRS XXth Congress*, 35, 12–23.

Angelov, P., Soares, E., 2019. Towards Explainable Deep Neural Networks (xDNN). *arXiv preprint arXiv:1912.02523*.

Ball, J. E., Anderson, D. T., Chan, C. S., 2017. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11(4), 042609.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 153–160.

Bergamasco, L., Saha, S., Bovolo, F., Bruzzone, L., 2019. Unsupervised change-detection based on convolutional-autoencoder feature extraction. *Image and Signal Processing for Remote Sensing XXV*, 11155, International Society for Optics and Photonics, 1115510.

Bovolo, F., Bruzzone, L., 2005. A detail-preserving scale-driven approach to change detection in multitemporal SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 43(12), 2963–2972.

Bovolo, F., Bruzzone, L., 2015. The time variable in data fusion: a change detection perspective. *IEEE Geoscience and Remote Sensing Magazine*, 3(3), 8–26.

Bruzzone, L., Prieto, D. F., 2000. A minimum-cost thresholding technique for unsupervised change detection. *International Journal of Remote Sensing*, 21(18), 3539–3544.

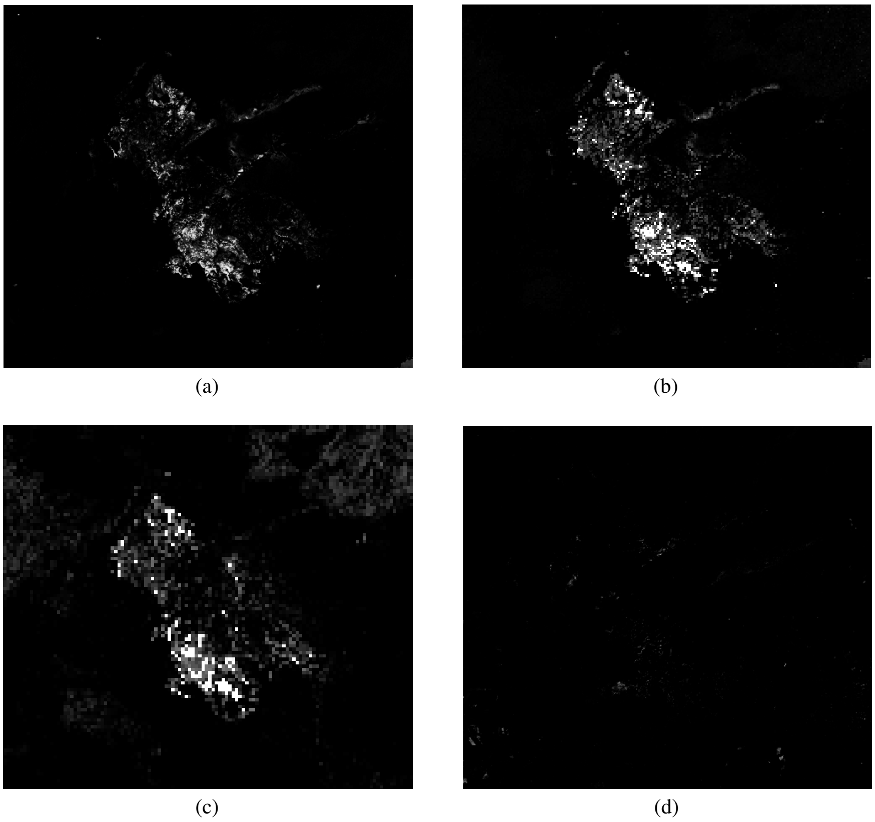
(a) (b) (c) (d)

Figure 5. Visualization of difference image generated by features: (a)-(c) 3 features retained by the method, (d) a feature discarded by the method.

Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE Geoscience and Remote Sensing Letters*, 6(4), 772–776.

Del Frate, F., Pacifici, F., Solimini, D., 2008. Monitoring urban land cover in Rome, Italy, and its changes by single-polarization multitemporal SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1(2), 87–97.

Guo, X., Liu, X., Zhu, E., Yin, J., 2017. Deep clustering with convolutional autoencoders. *International conference on neural information processing*, Springer, 373–382.

Huang, Z., Pan, Z., Lei, B., 2017. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. *Remote Sensing*, 9(9), 907.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature*, 521(7553), 436–444.

Liu, J., Gong, M., Qin, K., Zhang, P., 2016. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. *IEEE transactions on neural networks and learning systems*, 29(3), 545–559.

Ma, X., Geng, J., Wang, H., 2015. Hyperspectral image classification via contextual deep learning. *EURASIP Journal on Image and Video Processing*, 2015(1), 1–12.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. High-Resolution Semantic Labeling with Convolutional Neural Networks. *arXiv preprint arXiv:1611.01962*.

Otsu, N., 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1), 62–66.

Ozbulak, G., Aytar, Y., Ekenel, H. K., 2016. How transferable are cnn-based features for age and gender classification? *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 1–6.

Penatti, O. A., Nogueira, K., dos Santos, J. A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 44–51.

Roscher, R., Bohn, B., Duarte, M. F., Garcke, J., 2019. Explainable machine learning for scientific insights and discoveries. *arXiv preprint arXiv:1905.08883*.

Saha, S., Bovolo, F., Bruzzone, L., 2019a. Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Saha, S., Bovolo, F., Bruzzone, L., 2019b. Unsupervised multiple-change detection in vhr multisensor images via deep-learning based adaptation. *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 5033–5036.

Saha, S., Solano-Correa, Y. T., Bovolo, F., Bruzzone, L., 2019c. Unsupervised deep learning based change detection in sentinel-2 images. *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, IEEE, 1–4.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 881–893.

Xu, Y., Xiang, S., Huo, C., Pan, C., 2013. Change detection based on auto-encoder model for VHR images. *MIPPR 2013: Pattern Recognition and Computer Vision*, 8919, International Society for Optics and Photonics, 891902.

Xue, Q., Chuah, M. C., 2019. Explainable deep learning based medical diagnostic system. *Smart Health*, 13, 100068.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 22–40.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A., 2014. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 487–495.