DEEP FEATURE LEARNING FOR HYPERSPECTRAL IMAGE CLASSIFICATION AND LAND COVER ESTIMATION

Grigorios Tsagkatakis, Panagiotis Tsakalides

Institute of Computer Science, FORTH, Crete, Greece

Computer Science Department, University of Crete, Greece

ABSTRACT

The differences in spatial sampling between field measurements and remote-sensing imagery can hinder the exploitation of contemporary data. When the field-based sampling is higher than airborne and spaceborne imagery, each pixel is naturally associated with multiple pixels due to the multiplexing of the reflectances of different materials. To address this scale inconsistency, we propose the introduction of the multi-label classification framework where classifiers are trained to predict multiple labels per pixel. Furthermore, instead of relying on raw hyperspectral measurements for the classification process, we investigate the Stacked Sparse Autoencoders framework, an example of a deep learning network, for descriptive feature extraction. To validate the merits of the proposed scheme, we consider real data from the Hyperion instrument on-board the EO-1 and NYC land cover data from 2010.

Key words: Multi-label classification, feature learning, hyperspectral.

1. INTRODUCTION

Information contained in the electromagnetic spectrum is captured by Multispectral and Hyperspectral imaging devices which can provide key insights into the distribution of materials present in a scheme. Classification schemes exploit this information in order to assign individual or groups of pixel to the single most representative class, leveraging features extracted from labelled training examples. To fully exploit the available data, one must address the problem of *scale incompatibility* between field and remote sensing measurements. Whereas field-based measurements can be conducted at very fine resolutions, *e.g.*, meter scales, distance to the ground and motion of the moving platforms are directly responsible for the considerably lower spatial resolution of remote sensing imagery, especially spaceborne ones. This spatial scale incompatibility between field-based and satellite-based sampling inevitably introduces challenges in the exploitation of the acquired measurements.

In addition to the scale incompatibility, annotation of satellite data relies on the application of state-of-the-art classification methods that can leverage sufficient information from a limited number of training examples. Overall, the performance of the classification process primarily depends on two factors, namely the learning capacity of the classifier and the characteristics of the extracted features. The effects of the feature extraction process are particularly evident in computer vision tasks, where carefully designed, hand-crafted features, such as Scale Invariant Feature Transform (SIFT)[17] have shown great effectiveness in a variety of tasks. Despite their impressive performance, the main drawback of these descriptors is that significant human intervention is required during their design.

In remote sensing, similar features have been considered, including the Normalized Vegetation Difference Index (NDVI) and the Land Surface Temperature (LST). Such features are highly domain-specific and have limited generalization ability, especially when dealing with high spectral sampling rates, such as the ones in hyperspectral imaging. This motivates the need for efficient feature representations extracted automatically from raw data through Representation Learning [1], a set of techniques which aim to learn useful (i.e. discriminative, robust, smooth) representations of the input data for use in higher level tasks such as classification and recognition, minimizing the dependency of learning algorithms on feature engineering.

In this work, we first consider the problem of multi-label classification [23] where each satellite image pixel is annotated with multiple labels, encoding the different materials that can be mixed within a single pixel [13]. Furthermore, we seek “good representations” for satellite data under a real-world scenario by focusing on a particularly successful unsupervised feature learning approach by considering the deep learning framework of *Stacked Sparse Autoencoders* (SSAE), a type of artificial neural network which employs nonlinear codes and imposes sparsity constraints for representing the original data [12].

The rest of the paper is organized as follows. Section 2 gives a brief review of the recent endeavors in introducing

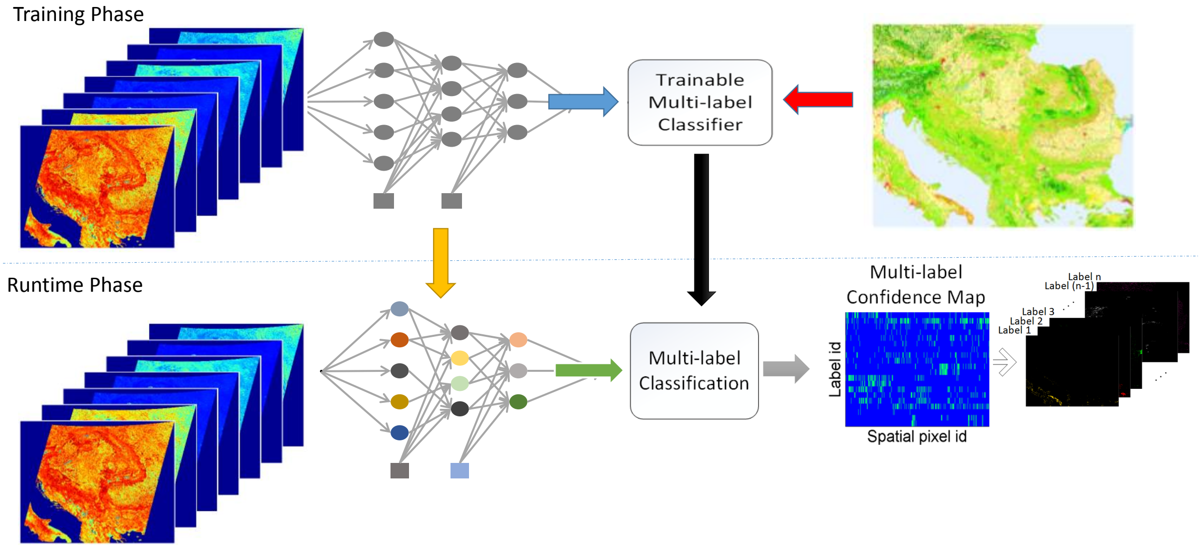

Figure 1: Block diagram of training (top part) and testing (bottom part) processes. During training, a deep learning network is first trained for feature extraction which are used for training a multi-label classifier.

deep learning approaches for the classification of remote sensing data. In Section 3 we outline the key theoretical components of SSAE and how they can be consider in the problem of multi-label classification. Section 4 provides an overview of the dataset along with experimental results, while the paper concludes in Section 5.

2. STATE-OF-THE-ART

Inspired by the human cognitive system which exhibits a hierarchical structure and learns in a layer-wise fashion, researchers have tried to incorporate depth into learning algorithms, which would allow to achieve function representation more compactly[3], and obtain increasingly more abstract representations. While it has been shown that one hidden layer can approximate a function to any level of precision, this approach becomes impractical due to the increase in the number of the required computational units [2].

Although theoretical results have been encouraging, in practice, training sufficiently deep architectures has been unattainable since gradient-based optimization methods starting from random initial weights tended to get fixated near poor local optima [14]. Deep Learning (DL) has gone through a revolution in the past decade by considering the strategy of greedy layer-wise unsupervised "pretraining" followed by supervised fine-tuning [10, 19].

DL has been recently considered for various problems in remote sensing data classification, including building detection from very high resolution multispectral data [27], classification and segmentation of Satellite Orthoimagery [16], and scene classification [31] among others. A classification framework composed of principal component analysis, deep convolutional neural networks and logistic regression was investigated in the context of spectralspatial classification of hyperspectral images [30]. Given the complexity of training a DL framework, the possibility of transferring models trained on everyday objects to remote sensing domain was investigated in [20]

In this work, we consider the framework of Autoencoders. Recently, several Autoencoders variants have been developed which introduce regularization in the latent space, including the denoising [28], the contractive [22], the saturating [8], and the sparse [19, 9] autoencoder. The technique of greedy layer-wise unsupervised "pretraining" has also be considered for Autoencoders [4]. Stacked Sparse Autoencoders (SSAE) have also been considered for the unsupervised spatio-spectral feature learning from hyperspectral imagery [25, 5] while other various such as Stacked Denoise Autoencoders [29] have also been explored.

3. FEATURE LEARNING FOR CLASSIFICATION

We consider training data consisting of deep learning features extracted from hyperspectral imagery acquired by the Hyperion instrument, and the corresponding land cover labels are utilized in order to build a multi-label mapping module. Once training is complete, a testing multispectral image can be annotated with multiple labels per pixel.

At a high-level, the basic modules of our system's pipeline are the following: (i) preprocessing and normalization of the features, (ii) feature-mapping using Stacked Sparse Autoencoders (SSAE) and (iii) multi-label classification based on the learned features. A visual description of the proposed scheme is given in Figure 1. In the following section, we present SSAE and how they can be applied in the concept of multi-label classification.

3.1. Stacked Sparse Autoencoders

Formally, a classical autoencoder is a deterministic feed-forward artificial neural network comprised of an input and an output layer of the same size with a hidden layer in between, which is trained with backpropagation [15] in a fully unsupervised manner, aiming to learn an approximation $\hat{\mathbf{x}}$ of the input which would be ideally more descriptive of the the raw input. The feature mapping that transforms an input pattern $\mathbf{x} \in \mathbb{R}^n$ into a hidden representation $\mathbf{h}$ (called code) of k neurons (units), is defined by the *encoder* function:

$$f(\boldsymbol{x}) = \boldsymbol{h} = \alpha_f(W_1\boldsymbol{x} + \boldsymbol{b_1}), \tag{1}$$

where $\alpha_f : \mathbb{R} \mapsto \mathbb{R}$ is the *activation function* applied component-wise to the input vector. The activation function is usually chosen to be nonlinear; examples include the logistic sigmoid and the hyperbolic tangent. The activation function is parametrized by a weight matrix $W_1 \in \mathbb{R}^{k\times n}$ with models the connections between the input and the hidden layer and a bias vector $\boldsymbol{b_1} \in \mathbb{R}^{k\times 1}$. The network output is then computed by mapping the resulting hidden representation $\boldsymbol{h}$ back into a reconstructed vector $\hat{\boldsymbol{x}} \in \mathbb{R}^{n\times 1}$ using a separate *decoder* function of the form:

$$g(f(\boldsymbol{x})) = \hat{\boldsymbol{x}} = \alpha_g(W_2\boldsymbol{h} + \boldsymbol{b_2}), \tag{2}$$

where α_g is the activation function, $W_2 \in \mathbb{R}^{n\times k}$ is the decoding matrix and $\boldsymbol{b_2} \in \mathbb{R}^n$ a vector of bias parameters which are learned from the hidden to the output layer.

The estimation of the parameters set $\theta = \{W_1, \boldsymbol{b_1}, W_2, \boldsymbol{b_2}\}$ of an autoencoder, is achieved through the minimization of the reconstruction error between the input and the output according to a specific loss function. Given the training set X, a typical loss function seeks to minimize the normalized sum of squares error, defining the following optimization objective:

$$J_{\text{AE}}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \|\frac{1}{2}x^{(i)} - \hat{x}^{(i)}\|^2 \tag{3}$$

where $\hat{x}$ is implicitly dependent on the parameter set θ and $\|\cdot\|$ is the Euclidean distance.

Sparse autoencoders are a special case of the traditional autoencoders, where the code is constrained to be sparse, *i.e.* only a small fraction of hidden units are activated by the inputs. Signal and model sparsity have had a profound impact on signal processing and machine learning due to their numerous advantages, such as robustness, model complexity, generative and discriminative capabilities among others[7, 26]. In order to induce the sparsity constraint, a sparsity constant ρ is selected and the average latent unit activation is enforced to be close to this value. This is achieved by introducing a Kullback-Leibler (KL) divergence regularization term, which measures the difference between Bernoulli distributions which encode the expected activation over the training set of hidden unit u ($\hat{\rho}_u$) and its target value (ρ) in our case:

$$\text{KL}(\rho||\hat{\rho}_u) = \rho \log\frac{\rho}{\hat{\rho}_u} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_u} \tag{4}$$

where $\hat{\rho}_u = \frac{1}{m}\sum_{i=1}^{m}\left[\alpha_u\left(x^{(i)}\right)\right], u = 1, \ldots, k$. The KL distance reaches its minimum of 0 when $\hat{\rho}_u = \rho$, and extends to infinity up as $\hat{\rho}_u$ increases, enforcing the $\hat{\rho}_u$ not to significantly deviate from the desired sparsity value ρ. All in all, the smaller the value of ρ, the sparser the representation would be. The regularized cost function of a sparse autoencoder constitutes of the reconstruction loss of a classical autoencoder with an additional regularization though a *sparsity promoting term* [18] given by:

$$J_{\text{spAE}}(\theta) = J_{\text{AE}}(\theta) + \beta\sum_{j=1}^{k} KL(\rho||\hat{\rho}_u)\ . \tag{5}$$

The hyper-parameter β determines the importance of the sparsity regularizer. Note that there have been also developed and other techniques to encourage sparsity in the representation [11] .

A particular set of weights is updated by calculating the partial derivatives of J_{spAE} and applying the backpropagation algorithm [15]. This way, the training typically converges to a minimum, hopefully a global one, after a small number of iterations. The minimization of the model parameters θ can be achieved by conventional optimization algorithms (*e.g.*, gradient descent), as well as with more sophisticated procedures, such as conjugate gradient and Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods to speed up convergence.

Deep learning is a special case of representation learning which aims at learning multiple hierarchical levels of representations, leading to more abstract features that are more beneficial in classification. Architectures with two or more hidden layers can be created by stacking single layer autoencoders on top of each other. Formally, one starts by training a sparse autoencoder with the raw data as input. Then the decoder layer is discarded so that the activations of the hidden units (layer 1 features) become the input for the second autoencoder, which in turn produces another representation (layer 2 features). This greedy layer-by-layer process keeps the previous layers fixed and ignores interactions with subsequent layers, thus dramatically reducing the search over the parameter space. We can formalize a stacked autoencoder according to:

$$h^{(L)} = f^{(L)}\left(\cdots f^{(2)}\left(f^{(1)}(x)\right)\right), \tag{6}$$

where $h^{(L)}$ denotes the representation learned by the top layer L.

Unsupervised pretraining [10] is a recently developed yet very influential protocol that helps to alleviate this problem by first training each layer independently in an unsupervised fashion and then performing a fine-tuning over the entire network based on the supervised classification error.

3.2. Multi-label classification

The features extracted by the stacked sparse autoencoders are then introduced for multi-label classification of multispectral pixels. In this work, we focus on a particular class of multi-label classifier, namely ensemble classifiers and more particular on the Ensemble of Classifier Chains (ECC) [21]. ECC has established itself as a powerful learning technique, based on the successful Classifier Chains (CC) model [21], which involves the training of m binary classifiers. In CC, the binary classifiers are linked along a "chain" so that each classifier is built upon the preceding ones. In particular, during the training phase, CC enhances the feature space of each link in the chain with binary features from ground-truth labeling. Since true labels are not known during testing, CC augments the feature vector by all prior binary predictions. Formally, the classification process begins with h_1 which determines $P(\lambda_1 \mid \boldsymbol{x})$, and propagates along the chain for every following classifier $h_2, \ldots, h_j$ predicting:

$$P(\lambda_j \mid \boldsymbol{x}, \lambda_1, \ldots, \lambda_{j-1}) \to \lambda_j \in \{0, 1\}, j = 2, \ldots, m \,. \tag{7}$$

The binary feature vector $(\lambda_1, \ldots, \lambda_m)$ represents the predicted label set of $\boldsymbol{x}$, $Z_{\boldsymbol{x}}$. Despite the incorporation of label information, the prediction accuracy is heavily dependent on the ordering of the labels, since only one direction of dependency between two labels is captured. To overcome this limitation, ECC extends this approach by constructing multiple CC classifiers with random permutations over the label space. Hence, each CC model is likely to be unique and able to give different multi-label predictions, while a good label order is not mandatory. More specifically, to obtain the output of ECC, a generic voting scheme is applied, where the sum of the predictions is calculated per label, and then a threshold t_s is applied to select the relevant labels, such that $\lambda_j \geq t_s$.

4. DATA DESCRIPTION AND EXPERIMENTAL RESULTS

We consider the Hyperion sensor aboard EO-1 with a spatial resolution of 30m^2, acquiring images at 242 spectral bands where we select only the 198 calibrated bands. We consider the area in New York city encoded as EO1H0130322010245110KF_SGS_01 by Hyperion from September 2, 2010. While global or European land cover datasets provide ground-truth data at relatively large spatial resolution, *e.g.* 30m^2, newer datasets offer a much higher spatial resolution. Such dataset do not consider widespread coverage as the process of labeling is extremely costly and time-consuming, yet they provide detailed maps of more specific geographic areas (*e.g.*, cities, forests, etc.). We consider a high resolution land cover dataset for New York City (NYC) of 2010 with a spatial resolution of 1m (3 feet) which been recently released[1]. The dataset annotates each spatial location with one of the following labels: (1) tree canopy, (2) grass/shrub, (3) bare earth, (4) water, (5) buildings, (6) roads, and (7) other paved surfaces.

The performance evaluation of multi-label classifiers is more complicated than conventional single-label learning, since an example may be partially correct. As a consequence, several metrics have been proposed for classification and ranking [32, 24]. In this work, we consider two representative error metrics, namely *Hamming Loss*, which measures the average number of locations where this is a discrepancy between predicted label and ground-truth label (lower is better) and *Averaged AUC*, the averaged Area-Under-the-Curve encodes the overall quality of performance, independently of individual threshold configurations (higher is better).

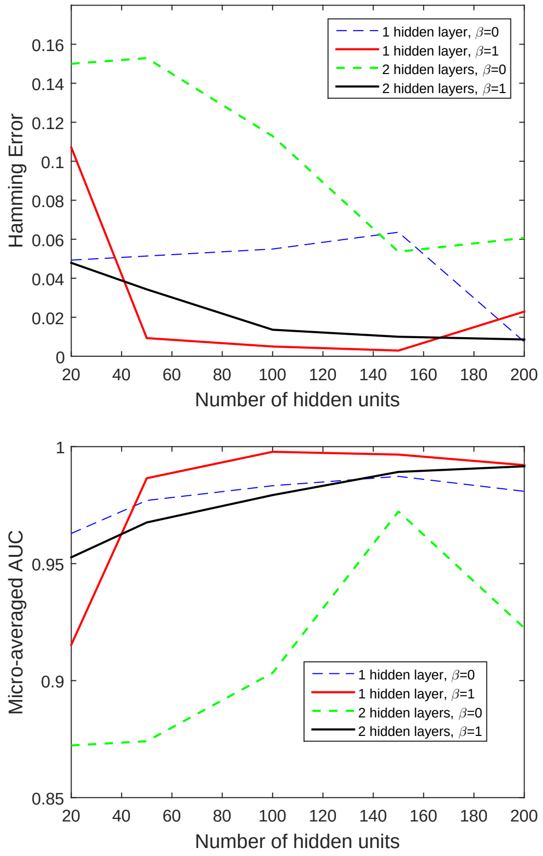

Figure 2: Hamming error (top) and Micro-averaged AUC (bottom) as a function of the number of hidden units for 1 and 2 hidden layers, as well as the regularization parameter for sparsity.

Figure 2 presents the classification performance of the ECC classifier based on features extracted by a SSAE architecture as a function of the number of hidden units considered in the hidden layers. These figures exam-

[1] https://nycopendata.socrata.com/Environment/Landcover-Raster-Data-2010-/9auy-76zt

ine two key parameters, (i) the number of hidden layers (depth) and (ii) the regularization parameter β with fixed sparsity target $\rho = 0.1$.

Regarding the number of hidden layer, the results indicate that if sufficient hidden units are considered, shallow architectures (1 layer) perform comparably to deep ones (2 layers), both in terms of Hamming error and Micro-averaged AUC. This situation is more pronounced when the sparsity promoting term (β) is active compared to inactive. Especially when the sparsity regularization is enabled, we observe that even a moderate number of hidden units and shallow architectures provide very good performance.

5. CONCLUSIONS

In this work we consider the case where relatively low hyperspectral images are available, where each pixel must be annotated with labels from a multi-label corpus. We investigate the potential of the recently developed deep learning paradigm, as an effective mechanism for extracting features that offer more abstract representations of the raw data. More specifically, we consider the paradigm of Stacked Sparse Autoencoders (SSAE) as an efficient mechanism feature extraction for multi-label classification. Experimental results suggest that although the deep of the network can aid in the classification process, the introduction of the sparsity constraints can have more dramatic gains in terms of performance.

ACKNOWLEDGMENTS

This work was partially funded by the PHySIS project (contract no. 640174) and the DEDALE project (contract no. 665044) within the H2020 Framework Program of the EC.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, Aug 2013.

[2] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[3] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, Universit De Montral, and Montral Qubec. Greedy layer-wise training of deep networks. In *In NIPS*. MIT Press, 2007.

[4] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7(6):2094–2107, June 2014.

[5] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7(6):2094–2107, 2014.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[7] Konstantina Fotiadou, Grigorios Tsagkatakis, and Panagiotis Tsakalides. Low light image enhancement via sparse representations. In *Image Analysis and Recognition*, pages 84–93. Springer International Publishing, 2014.

[8] Rostislav Goroshin and Yann LeCun. Saturating auto-encoder. *CoRR*, abs/1301.3577, 2013.

[9] Ian Goodfellow, Honglak Lee, Quoc V. Le, Andrew Saxe, and Andrew Y. Ng. Measuring invariances in deep networks. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 646–654. Curran Associates, Inc., 2009.

[10] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[11] K. Kavukcuoglu, M.A. Ranzato, R. Fergus, and Yann LeCun. Learning invariant features through topographic filter maps. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1605–1612, June 2009.

[12] Konstantinos Karalas, Grigorios Tsagkatakis, Michalis Zervakis, and Panagiotis Tsakalides. Deep learning for multi-label land cover classification. In *SPIE Remote Sensing*, pages 96430Q–96430Q. International Society for Optics and Photonics, 2015.

[13] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides. Land classification using remotely sensed data: Going multilabel. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3548–3563, June 2016.

[14] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *J. Mach. Learn. Res.*, 10:1–40, June 2009.

[15] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In Grégoire Montavon, GeneviveB. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 9–48. Springer Berlin Heidelberg, 2012.

[16] Martin Längkvist, Andrey Kiselev, Marjan Alirezaie, and Amy Loutfi. Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing*, 8(4):329, 2016.

[17] DavidG. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] Andrew Ng. Sparse autoencoder. *CS294A Lecture notes*, 72, 2011.

[19] Christopher Poultney, Sumit Chopra, and Yann Lecun. Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems (NIPS 2006*. MIT Press, 2006.

[20] Otavio Penatti, Keiller Nogueira, and Jefersson Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 44–51, 2015.

[21] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. volume 85, pages 335–359, 2011.

[22] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contracting auto-encoders: Explicit invariance during feature extraction. In *In Proceedings of the Twenty-eight International Conference on Machine Learning (ICML11*, 2011.

[23] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *Int. J. of Data Warehousing and Mining*, 3(3):1–13, 2007.

[24] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.

[25] Chao Tao, Hongbo Pan, Yansheng Li, and Zhengrou Zou. Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *Geoscience and Remote Sensing Letters, IEEE*, 12(12):2438–2442, 2015.

[26] Grigorios Tsagkatakis and Andreas Savakis. Sparse representations and distance learning for attribute based category recognition. In *Trends and Topics in Computer Vision*, pages 29–42. Springer Berlin Heidelberg, 2012.

[27] M Vakalopoulou, Konstantinos Karantzalos, Nikos Komodakis, and Nikos Paragios. Building detection in very high resolution multispectral data with deep learning features. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, pages 1873–1876. IEEE, 2015.

[28] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM.

[29] Chen Xing, Li Ma, and Xiaoquan Yang. Stacked denoise autoencoder based feature extraction and classification for hyperspectral images. *Journal of Sensors*, 2016, 2015.

[30] Jun Yue, Wenzhi Zhao, Shanjun Mao, and Hui Liu. Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sensing Letters*, 6(6):468–477, 2015.

[31] Yanfei Zhong, Feng Fei, and Liangpei Zhang. Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *Journal of Applied Remote Sensing*, 10(2):025006–025006, 2016.

[32] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, Aug 2014.