Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
13/1 (2020), 9-15. DOI: http://dx:doi:org/10:21609/jiki:v13i1:761

FULLY CONVOLUTIONAL VARIATIONAL AUTOENCODER FOR FEATURE EXTRACTION OF FIRE DETECTION SYSTEM

Herminarto Nugroho[1], Meredita Susanty[2], Ade Irawan[3], Muhammad Koyimatu[4], Ariana Yunita[5]

[1] Departement of Electrical Engineering, Faculty of Industrial Technology, Pertamina University, Jalan Teuku Nyak Arief, Simprug, Kebayoran Lama, Jakarta Selatan, 12220, Indonesia.
[2,3,4,5] Departement of Computer Science, Faculty of Science and Computer, Pertamina University, Jalan Teuku Nyak Arief, Simprug, Kebayoran Lama, Jakarta Selatan, 12220, Indonesia.

E-mail: herminarto.nugroho@universitaspertamina.ac.id

Abstract

This paper proposes a fully convolutional variational autoencoder (VAE) for features extraction from a large-scale dataset of fire images. The dataset will be used to train the deep learning algorithm to detect fire and smoke. The features extraction is used to tackle the curse of dimensionality, which is the common issue in training deep learning with huge datasets. Features extraction aims to reduce the dimension of the dataset significantly without losing too much essential information. Variational autoencoders (VAEs) are powerfull generative model, which can be used for dimension reduction. VAEs work better than any other methods available for this purpose because they can explore variations on the data in a specific direction.

Keywords: *variational autoencoder, feature extraction, deep learning, computer vision, fire detection system*

Abstrak

Artikel ini membahas mengenai *fully convolutional variational autoencoder* (VAE) untuk ekstraksi fitur dari dataset gambar api. Dataset akan digunakan untuk melatih algoritma *deep learning* untuk mendeteksi api dan asap. Ekstraksi fitur digunakan untuk mengatasi *the curse of dimensionality*, yang merupakan masalah umum dalam *training deep learning* dengan ukuran dataset yang sangat besar. Ekstraksi fitur bertujuan untuk mengurangi dimensi dataset secara signifikan tanpa kehilangan terlalu banyak informasi penting. *Variational autoencoder* (VAE) adalah model generatif yang kuat, yang dapat digunakan untuk pengurangan dimensi. VAE bekerja lebih baik daripada metode lain yang tersedia untuk tujuan ini karena mereka dapat mengeksplorasi variasi pada data.

Kata Kunci: *variational autoencoder, feature extraction, deep learning, computer vision, fire detection system*

1. Introduction

Fire detection is commonly performed visually by using ultraviolet (UV) camera [1], infrared (IR) camera [2], or visible light camera [3]. UV-based and IR-based fire detection has high sensitivity and fast response, yet prone to disturbance from other UV and IR source light [4] [5]. Hence, this paper focuses on the fire detection system based on the visible light camera which uses a charged-coupled device (CCD) sensor. CCD sensor records a glimpse of fire in the form of video or static images as the data. Computer vision techniques then preprocess the data prior to data training. The data training exploits Deep Learning algorithm to detect whether or not the flame exists in the captured video or images. The deep learning algorithm has been implemented to solve many complex problems [6] [7]. The algorithm can increase the accuracy of detection from any kind of fire and smoke in the captured videos or images [8]. However, the algorithm needs a huge number of datasets in order to obtain high accuracy detection and hence costs computationally expensive. Therefore, it is desirable to extract only the important features of the captured videos or images, such that the dimension of the datasets can be reduced while the most of the information in the data is still preserved.

One of the methods for fire detection using

CCD sensor is by using a rule-based generic color model, which uses *YCbCr* color space to separate the illuminance from the chominance [9]. The using of *YCbCr* is indeed more effective than using *RGB* color space to separate illuminance. This method produces high fire detection accuracy and reasonable false alarm rate. However, this method only relies on the color detection of fire. Other important features of fire, for example smoke, and other color of fire, i.e., blue fire, cannot be detected quite well.

The proposed method in this paper aims to detect all important features of fire. In order to extract all the important features of fire, many techniques have been developed for the purpose of feature extraction, such as auto-encoder [10], Isomap [11], Nonlinear Dimensionality Reduction (NLDR) [11], Multifactor Dimensionality Reduction (MDR) [13], and Principal Component Analysis (PCA) [14].

Principal Component Analysis (PCA) has been widely used for feature extraction. However, PCA only attempts to discover a lower dimensional hyper lane which describes the original dataset. In other word, PCA only tries to learn linear manifolds of datasets. This will result in the feature extracted can loose many information. On the other side, a neural network-based feature extraction, for example auto encoder, is capable to learn nonlinear relationships or manifolds from datasets.

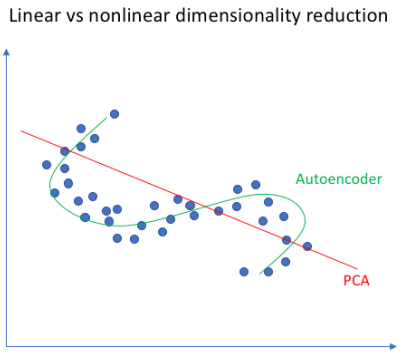

Figure 1. Comparison between PCA and Autoencoder [15]

Auto-encoder has been used widely for feature extraction in images datasets because of its robustness to the noise and disturbance in the images [14]. As one of the examples, the stacked denoising auto-encoder has been implemented for feature extraction and classification of hyperspectral images [15]. Furthermore, the introduction of the stacked convolutional auto-encoders has proven to significantly reduce the computation cost for feature extraction process [16].

Auto-encoder consists of a pair of two connected networks: an encoder and a decoder. An encoder takes an input and then converts it into a hidden representation which has significantly smaller dimension compared to the input vector. This hidden representation refers to the features which are extracted from the given input. It is then mapped back by a decoder to obtain the output of the network which reconstruct or generate the given input with high probability [17]. The auto-encoder output will not exactly reconstruct the input because of the existence of the reconstruction error. The reconstruction error function is usually either the mean-squared error [18] or the cross-entropy [19] which penalizes the network for creating outputs different from the input. It depends on the dimension of the hidden representation or the extracted features. The smaller the dimension of the hidden representation, the bigger the reconstruction error becomes. This create the trade-off between the dimension of the features and the information loss. It is desirable that the dimension of the features can be minimized while most of the information in the data is still retained.

Standard plain auto-encoder indeed is able to generate a dense representation and reconstruct the input well. However, it is limited to a certain implementation only. The fundamental problem with the standard auto-encoder is that the latent space (the dense hidden representation/decoded vectors) and the encoded vectors may not be continuous, or even though they are continuous, they may be difficult to interpolate [20]. For example, auto-encoder works well for replicating the MNIST [21] or Fashion-MNIST dataset [22]. This is caused by the characteristic of the image datasets from MNIST and Fashion-MNIST is relatively simple and easy to distinguish between background and foreground. However, when dealing with more complex image datasets and the generative model, i.e., generating variations on the input dataset from the latent space, standard auto-encoder will not work well because of the discontinuities in the latent space [23]. As we know, fire has no standard distinguishable form, has many colors, and sometimes is covered by smoke, creating difficulties to extract useful features using plain auto-encoder. For this reason, this paper proposes feature extraction method using variational auto-encoder (VAE).

The organization of this paper is as follows. The fundamental concept of Variational Auto-Encoder (VAE) is introduced in Section 2. Section 3 presents the proposed architecture of the Variational Auto-Encoder used for feature extraction from fire images, which is Fully Convolutional Auto Encoder. The implementation results of the proposed Fully Convolutional Auto Encoder for fire feature extraction is shown in Section 4. In Section 5, the implications of the

proposed method are presented. Finally, the conclusion is given in Section 6.

2. Variational Autoencoder (VAE)

Variational Autoencoder network is a pair of two connected network – a network that takes in an input and produce smaller representation (encoder), and a network that convert back the smaller representation to the original input (decoder) that have continuous latent space, easy random sampling and interpolation because its encoder outputting two vectors – a vector of means (μ) and a vector of standard deviation/variance (σ) as illustrated in Figure 2.

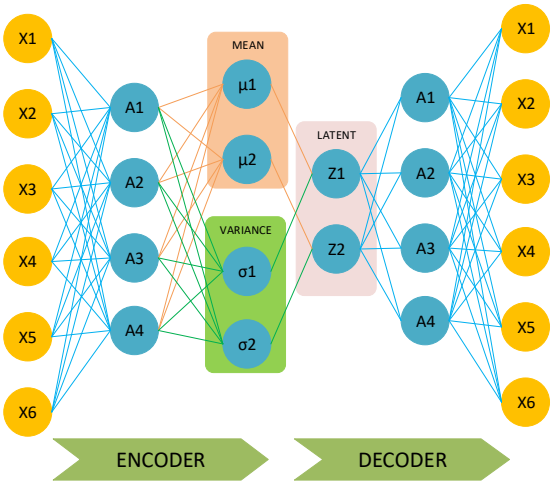

Figure 2. Basic structure of Variational Autoencoder (VAE)

As the encoding has far less units than the input, the challenge is getting the model to learn a meaningful and generalizable latent space. VAE encoder describes a probability distribution for each latent attribute. The two vectors form the parameters of a vector of random variables of length n, with the i th element of μ and σ being the mean and standard deviation of the i-th random variable, X_i, from which we sample, to obtain the sampled encoding. For the same input, although the mean and standard deviations remain the same, the actual encoding will somewhat vary on every single pass simply due to sampling. The mean vector controls where the encoding of an input should be centered around, while the standard deviation controls the "area", how much from the mean the encoding can vary. The Kullback–Leibler divergence that is used in loss function as a regularizer allowing smooth interpolation and enabling the construction of new samples.

The decoder network then subsequently takes random sample from each latent state distribution to generate a vector as input for our decoder model and attempts to recreate the original input. Backpropagation which is usually used to calculate the relationship of each parameter in the network with respect to the final output loss, cannot be used for random sampling process, thus reparametrizes is used instead. Using reparameterization, parameter of the distribution is optimized while still maintaining the ability to randomly sample from that distribution.

3. Proposed Fully Convolutional Autoencoder (FCAE) Architecture

The architecture of our network is summarized in Figure 3. It contains three main structure: the encoder, the bottleneck, and the decoder. Since the architecture proposed in this paper is fully convolutional variational autoencoder, all layers in the network architecture is convolutional layers. The input of the network is the given fire image taken randomly from the fire image datasets, while the output of the network is the reconstruction image from the given input image. Both the imput and output images are RGB images.

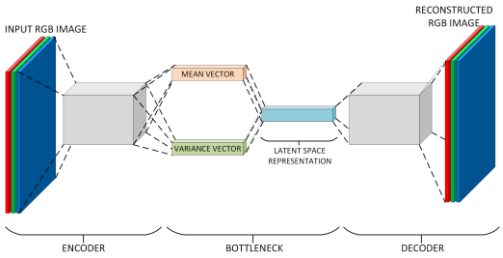

Figure 3. The structure of the proposed Fully Convolutional Autoencoder for feature extraction of fire image datasets.

3.1. The Encoder Structure

The encoder structure is a sequential network consisting of four convolutional layers with ReLU non-linearity for each respected convolutional layer. The kernel size used in the convolutional layer is 4, with stride 1 and no padding. Figure 3 shows the illustration of the encoder structure.

Consider the input image has $X \times Y$ pixels dimension. It means that $n_w^{[0]} = X$ and $n_h^{[0]} = Y$, where $n_w^{[0]}$ and $n_h^{[0]}$ specify the width and height dimension of the input layer. The formula to find the width and height dimension of the next layer is

$$n_w^{[l]} = \left\lfloor \frac{n_w^{[l-1]} + 2p - f}{s} + 1 \right\rfloor, \quad (1)$$

$$n_h^{[l]} = \left\lfloor \frac{n_h^{[l-1]} + 2p - f}{s} + 1 \right\rfloor, \quad (2)$$

where p defines the padding size, f specifies the

filter/kernel size, and s specifies the stride size. From equation (1-2) we know that because of the zero-padding used, the resulting output of this encoder sequential network has lower dimention compared to the input image.

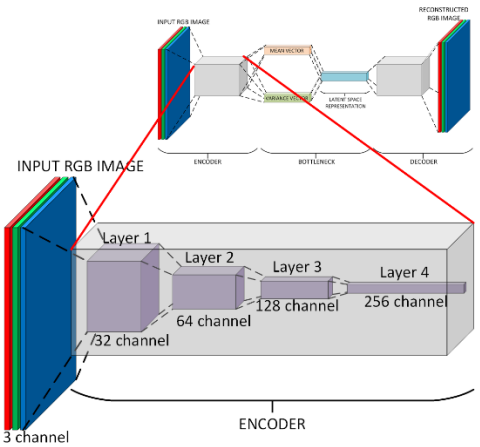

Figure 4. The structure inside encoder of the proposed Fully Convolutional Autoencoder for feature extraction of fire image datasets.

3.2. The Bottleneck Structure

This bottleneck structure is what unique from variational auto-encoder (VAE), compared to plain auto-encoder. While in plain auto-encoder the decoder will give one output of latent vector, VAE gives two outputs of vector means and vector variance with the same dimension. In the fully convolutional architecture, both of mean and variance vector are convolutional layers. The dimension of mean and variance vector specifies the dimension of the feature points extracted from the given image. To find the mean and variance vector, most literature use the Kullback–Leibler divergence (KL divergence [24]) as the loss function. Minimizing KL divergence means optimizing the probability distribution parameter to closely resemble the target distribution. For VAE, the KL divergence loss function is shown in the following equation:

$$\sum_{i=1}^{n} \sigma_i^2 + \mu_i^2 - \log(\sigma_i) - 1, \qquad (3)$$

where σ specifies the variance, and μ specifies the mean vector.

3.3. The Decoder Structure

The decoder structure is a transpose of the encoder structure. It is also a sequential network consisting of four convolutional layers with ReLU non-linearity for each respected convolutional layer. The same stride and padding used in the encoder is used in the decoder as well. However, we must carefully set the kernel size for each convolutional layer using equation (1-2) to reconstruct the images with the same pixel size compared to the input image. Figure 4 shows the illustration of the decoder structure.

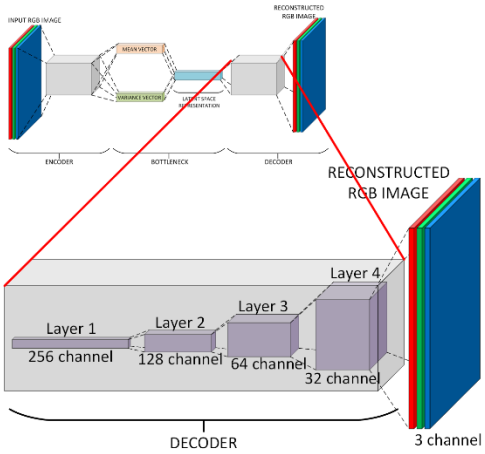

Figure 5. The structure inside decoder of the proposed Fully Convolutional Autoencoder for feature extraction of fire image datasets.

4. Feature Extraction: Result and Analysis

Unlike face or any landscape background, the nature of fire makes it is hard to extract the feature of fire from an image or video. Fire has many distinctive forms and colors. Sometimes, the fire is covered by smoke which makes it more difficult to extract its features. In this section, we present the result of the feature extraction of a dataset containing 10793 RGB images that mostly contain fire, but has 10% outlier images, i.e. images which do not contain fire. The image resolution is 480×480 pixels. Thus, these raw images have initially 3×230400 feature points each. It is a relatively large feature dimension compared to a small resolution image. This fact shows the importance of feature reduction to save computation power.

4.1. Feature Extraction from the Fire Images Dataset

Figure 6 shows the extracted feature points taken from randomly chosen images from the dataset. Each box in Figure 6 (a) consists of feature points from 64 randomly chosen images (8×8 image matrix) for a certain VAE training iteration. From the comparison between each feature from each iteration, we can observe that the VAE algorithm tries to learn which important features should be saved and which information may be omitted. In the iteration 12, we can observe that the extracted features of every image are different with each other. The resulting extracted features from this process can be used as a substitution for the initial

image for deep learning-based fire detector. However, we should confirm first that we can reconstruct the initial images from these features. If we can distinguish the image with fire or not from the reconstruted images, then it means the features can be used as the substitution for the initial images in the dataset.

4.2. Reconstruction Result from the Extracted Features

Figure 7 shows the reconstruction result using the features obtained for each iteration in Figure 6. As the VAE learn to extract important features, the reconstructed images become clearer and more distinguishable. We can observe from the reconstruction result using features from iteration 12, we can distinguish image with fire and image without fire. This means, the feature points in the iteration 12 contain enough important information. Therefore, they can be used to substitute initial images from the dataset.

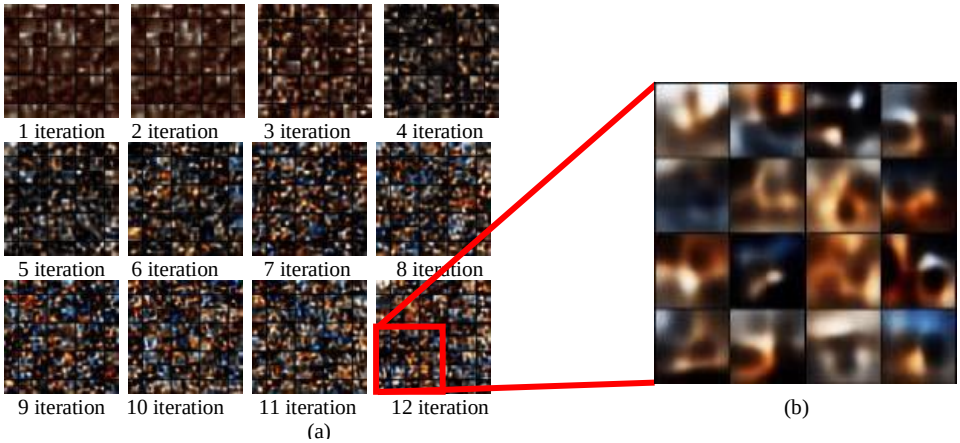

Figure 6. The result of feature extraction from 64 randomly chosen images from the image datasets. (a) The extracted features from each iteration process, and (b) the enlarged version of the features extracted from the image datasets.

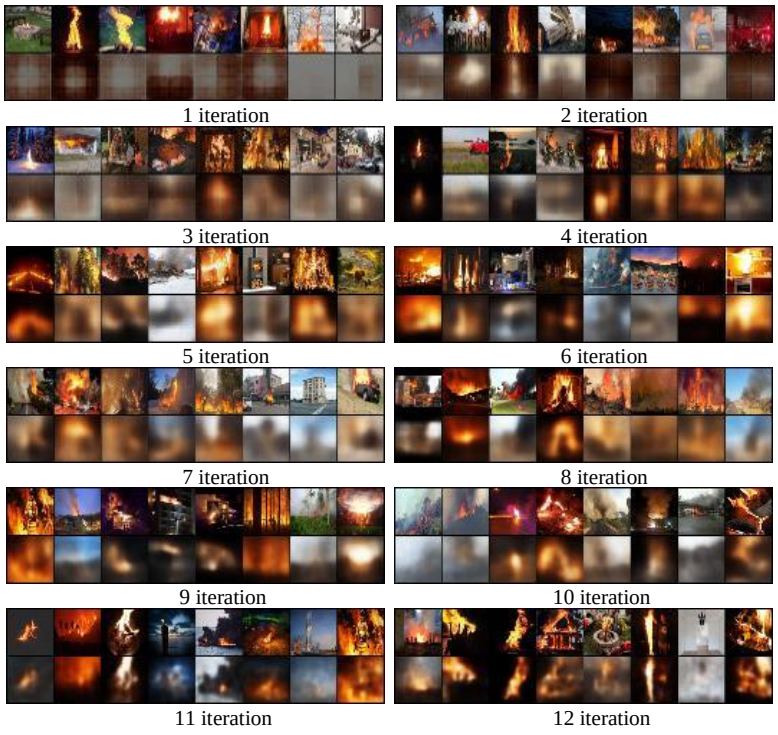

Figure 7. Reconstruction result from the extracted features using Fully Convolutional Autoencoder with different feature points extracted from each iteration

5. Implications of the Proposed Method

This method can be implemented for fire detection in buildings, as an addition for the already existing fire detection sensor. CCTV are already common to be placed in the surrounding of the building, and therefore can be used as a fire detection system.

6. Conclusion and Future Research

Fully Convolutional Variational Autoencoder (VAE) is suitable to extract features from a given fire images dataset. Even though the nature of fire makes hard to extract the feature of fire from an image, Fully Convolutional VAE can actually extract enough important features. The resulting extracted features then can be reconstructed and still can be distinguished between images which contain fire or not. From this reconstruction images, we can determine the suitable latent vector which results in the smallest feature points without losing too much important information. This suitable latent vector then can be used to substitute the initial images in the dataset. This latent vector by nature has significantly smaller dimension compared to the initial image. For future work, it is interesting to compare this algorithm to another feature extraction method such as Isomap, nonlinear dimensionality reduction (NLDR), or multifactor dimensionality reduction (MDR).

Acknowledgements

This research is funded by PT. Pertamina (Persero) via Universitas Pertamina selective research grant (UPSKILLING) 2018 with grant number 002/UP-WR3.1.1/SK/III/2018 date March 20, 2018. All source code and images of this research is available on https://github.com/herminarto/DLCV

References

[1] J. Sun, G. Guo and X. Zhang, "Research on UV Flame Detector," in Instrumentation and Measurement, Computer, Communication and Control (IMCCC), 2014 Fourth International Conference on. IEEE, 2014.

[2] L. Yang, Z. Guo-Sheng, L. Li-Kun and Z. Chong, "Research on the stability of IUR76-I/IUR76-II test systems for flame detectors and related national standards," in Test and Measurement, 2009. ICTM'09. International Conference on, 2009.

[3] J. Choi and J. Y. Choi, "Patch-based fire detection with online outlier learning," in Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on, 2015.

[4] G. Monitors, "How to Select a Flame Detector.," 10 July 2018. [Online]. Available: http://www.gmigasandflame.com/article_how_to_select_a_flame_detector.html

[5] Spectrex, "Spectrex Flame Detector Range," 10 July 2018. [Online]. Available: http://www.technoswitch.co.za/wp-content/uploads/2017/01/BR-Spectrex-170130.pdf

[6] H. Nugroho, "Tuning of Optical Beamforming Networks: A Deep Learning Approach," TU Delft, Delft, 2015.

[7] H. Nugroho, W. K. Wibowo, A. R. Annisa and H. M. Rosalinda, "Deep Learning for Tuning Optical Beamforming Networks," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 16, no. 4, 2018.

[8] X. Wu, X. Lu and H. Leung, "An adaptive threshold deep learning method for fire and smoke detection," in Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on, 2017.

[9] J. Masci, U. Meier, D. Ciresan and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," International Conference on Artificial Neural Networks, pp. 52--59, 2011.

[10] T. Celik and H. Demirel, "Fire detection in video sequences using a generic color model," Fire Safety Journal, vol. 44, no. 2, pp. 147--158, 2009.

[11] J. Masci, U. Meier, D. Ciresan and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," International Conference on Artificial Neural Networks, pp. 52--59, 2011.

[12] D. Lunga, S. Prasad, M. M. Crawford and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," IEEE Signal Processing Magazine, vol. 31, no. 1, pp. 55--66, 2014.

[13] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," SIAM journal on scientific computing, vol. 26, no. 1, pp. 313--338, 2004.

[14] J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, N. Barney and B. C. White, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," Journal of theoretical biology, vol. 241, no. 2, pp. 252--261, 2006.

[15] T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Phil. Trans. R. Soc. A, vol. 374, no. 2065, 2016.

[16] P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in Proceedings of the 25th international conference on Machine learning, 2008.

[17] C. Xing, L. Ma and X. Yang, "Stacked denoise autoencoder based feature extraction and classification for hyperspectral images," Journal of Sensors, 2016.

[18] J. Masci, U. Meier, D. Ciresan and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in International Conference on Artificial Neural Networks, 2011.

[19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," Journal of machine learning research, vol. 11, no. Dec, pp. 3371--3408, 2010.

[20] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks," in Advances in neural information processing systems, 2007.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," science, vol. 313, no. 5786, pp. 504--507, 2006.

[22] C. Doersch, "Tutorial on variational autoencoders," arXiv preprint arXiv:1606.05908, 2016.

[23] Y. LeCun, "The MNIST database of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[24] H. Xiao, K. Rasul and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.

[25] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens and L. Carin, "Variational autoencoder for deep learning of images, labels and captions".

[26] J. M. Joyce, "Kullback-leibler divergence," in International encyclopedia of statistical science, Springer, 2011, pp. 720--722