AutoEncoders for Denoising and Classification Applications

Boris A. Assanovich
Yanka Kupala State University of Grodno
Grodno, Republic of Belarus
bas@grsu.by

***Abstract*—Several structures of autoencoders used for the efficient data coding with unsupervised learning and applied to solving the tasks of classification and removing the internal noise from data used in problems of biometric and emotional recognition have been analyzed in this paper. Smile type recognition and biometric identification experiments using the transformed features from UvA-NEMO Smile Database and Caltech Faces datasets showed the possibility of improving the classification accuracy by 10%**

***Keywords*—stacked autoencoder, denoising, classification accuracy, unsupervised learning, smile type recognition, biometric identification**

I. Introduction

Learning of data representations with little or no supervision is a key challenge in machine learning (ML). Last time, there has been an increasing interest in the use of autoencoders to solve many theoretical and applied data processing tasks based on principals of ML, especially considering how to learn a mapping from high-dimensional observations to a lower-dimensional representation space [1]. However, recent advances in the use of auto-encoders are largely based on the seminal paper [2], which served as the beginning of the development of new algorithms for data processing in ML.

Autoencoder is a special architecture of artificial neural networks that allows the use of unsupervised learning using the method with back propagation of error. Generally speaking, an autoencoder is a direct distribution network without feedback, most similar to a perceptron and containing an input layer, one or several intermediate layers and an output layer. The goal of an autoencoder is to learn a representation (encoding) for a dataset, typically for dimensionality reduction, by training the network and ignore the "noise" signal. Autoencoders are used to solve many applied problems, from face recognition to obtaining the semantic meaning of image and language structures.

In this paper, we consider the use of autoencoders for solving applied problems of classification and removing the internal noise from data used in biometric and emotional recognition. Further, in the first section, the formalization of the construction of autoencoders types are considered, in the second section, the results of experiments using video and image processing datasets from the UvA-NEMO Smile Database and Caltech Faces database are presented. In conclusion, the evaluation of experiments is given, further aspects for the use of autoencoders are determined, which completes the paper.

II. Autoencoder Structure and Types

There are several types of autoencoders. *Sparse Autoencoder* has a dimension of the hidden layer that is greater than the input. It consists of two parts: coder (encoder) G and decoder F as depicted in Figure 1.

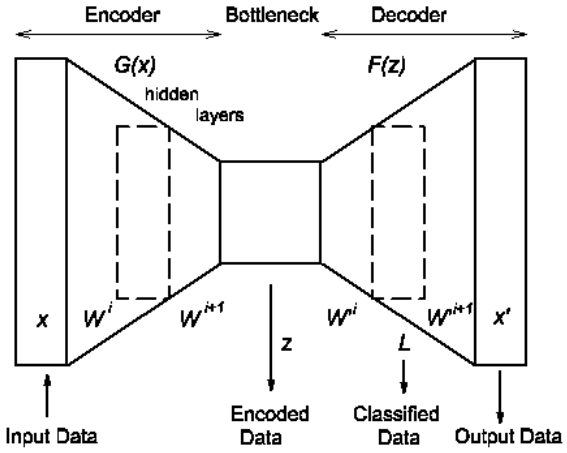

Figure 1. Autoencoder Structure

The encoder translates the input signal into its representation (code): $z = G(x)$, and the decoder restores the signal by its code: $x' = F(z)$. Moreover, the transformation functions F and G contain activation function, weights and biases of trained artificial neural networks [2]. By changing the mappings F, G, the autoencoder tends to learn the identity function $x = F(G(x))$, minimizing kind of error based on some functional $L(x, F(G(x')))$.

Let consider that a vector $x \in R$ connected to the input of an autoencoder. Then the encoder maps the vector x to another vector $z \in R$ as follows $z = h^i(W^i x + b^i)$,

where the superscript i indicates the i-th layer. Then h^i is a transfer function for the encoder, $W^i \in R$ is a weight matrix, and $b^i \in R$ is a bias vector. Hence, the decoder maps the encoded representation z back into an estimate of the original input vector x, as follows: $x' = h^{i+1}(W^{i+1}z + b^{i+1})$, where the superscript $i+1$ represents the $i+1$ layer. Then a transfer function h^{i+1} for the decoder has a factor $W^{i+1} \in R$ that is a weight matrix, and $b^{i+1} \in R$ is a bias vector correspondently. Hence, if the encoder has only 2 layers then the expression for the transfer function can be represented as $x' = h^2(W^2 z + b^2)$.

Usually autoencoders are limited in the dimension of the code (it is smaller than the dimension of the signal). The input signal is restored with errors due to coding losses, but in order to minimize them, the network is forced to learn to select the most important features.

To exclude the process of overfitting in ML the sparse autoencoders are used. It is done usually by adding a regularizer to the cost function. This regularizer is a function of the average output activation value of a neuron and encourages the autoencoder to learn a representation, where each neuron in the hidden layer "fires" to a small number of training examples. That makes to respond of neurons to some feature that is only present in a small subset of training examples. Then, for a sparse autoencoder a cost function consists of 3 terms: mean squared error term, regularization term and sparsity regularization term [3]. These parameters are usually used to optimize its unsupervised training.

Another type of autoencoder is a denoising autoencoder (DAE). It gets trained to use a hidden layer to reconstruct a particular model based on its inputs. DAE take a partially corrupted input and is trained to recover the original *undistorted input by minimizing* the loss function between the output node and the damaged input.

Recently, such structure as stacked autoencoder (SAE) has become popular. It is a neural network that includes several layers of sparse autoencoders where output of each hidden layer is connected to the input of the successive hidden layer. In this case the hidden layers are trained by an unsupervised algorithm and then fine-tuned by a supervised method.

A stacked denoising autoencoder (SDA) can be simply represented as several DAE that perform denoising. A key function of SDAs is unsupervised layer by layer pre-training. When a network is being trained, it generates a model, and measures the distance between that model and the benchmark through a loss function. SDA attempts to minimize the loss function involve resampling the damaged inputs and re-reconstructing the data, until it finds those inputs which bring its model closest to what it has been told is true.

Last time, of particular interest is the variational autoencoder, in which instead of mapping an input to fixed vector, input is mapped to a distribution. However, its description is beyond the scope of this study.

III. Application of SDA for Noise Reduction and Classification

In our opinion, the conversion of input signals and the analysis of transformed features received by an autoencoder is more effective approach for data processing, as it allows to reduce the data size in latent space and complexity of the entire overall system, as well as use a more accurate data separation.

We developed the FaceAnalyzer Platform (FAP) [4] based on OpenFace tool [5] to carry out the study of human characteristics and perform the analysis, recognition and verification of human biometric data, elements of his emotions by capturing images and video, features extraction and data processing as shown in Figure 2.

In [6], we have made an analysis of two smiles types temporal characteristics with the use of lip corner displacement, and perform their classification exploiting a well-known k-nearest neighbors (k-NN) algorithm based on the intensity of a person's face Action Units (AU), which describe the physiological dynamics of a human smile. Generally, smiles can be composed of three non-overlapping phases; the onset (neutral to expressive), apex, and offset (expressive to neutral), respectively. We made an attempt to use the detection of these three phases to perform the classification between genuine and posed smiles.

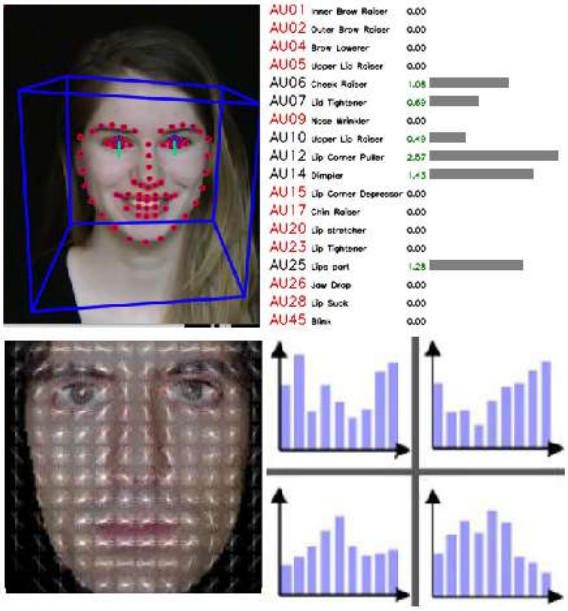

Figure 2. AUs and HOG-data from FAP

The analysis has been performed with FAP that perform 68 Facial Landmark detection and tracking with the help of pre-trained models based on Conditional Local

Neural Fields [5] (CLNF). To perform a classification, the timing smile signals were segmented into three phases, describing the state of a smiling person's face, as it was described in the work [6]. In each frame of analyzed smile, the intensities of the AU measurements were made, averaged over all frames, and a vector with nonzero intensity values $x = [x_1 \dots x_n]$ was formed for learning and classification purposes. The k-NN classification algorithm for the different values of k has been applied in the experimental setup. The results with a vector length of $n = 17$ elements showed that the classification accuracy is 60%. Next, we increased the dimension of vector by separating and averaging the AU intensity values over all 3 phases (Onset Apex and Offset) and used the increased feature vectors with 3x17=51 elements. The classification results with the use of k-NN algorithm have been slightly improved with weakly dependence on k number.

In further studies, we have designed SAE that consisted of 2 sparse autoencoders as depicted in Figure 3. Their convolutional layers were used to apply data de-noising, reduce features dimension and perform classification. The dimension of input vector (hiddenSize) for the first autoencoder was 51 elements which was pre-trained by Matlab with parameters: L2WeightRegularization', 0.001, 'SparsityRegularization', 4, 'SparsityProportion', 0.05, 'DecoderTransferFunction','purelin'. For the second autoencoder we have used similar values of parameters and dimension of 17 elements. The output layer of the second autoencoder had 8 elements, to which the soft-max decision function has been attached to perform a binary classification (upper part of Figure 3). The classification accuracy with the use of SAE achieved 70%.

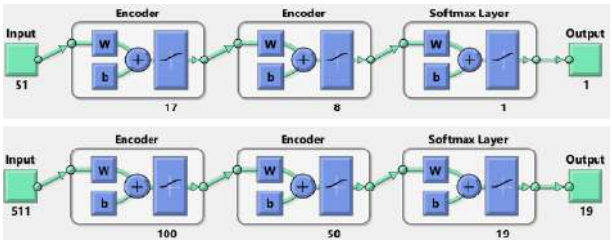

Figure 3. Stacked Autoencoders

The next experiment has been conducted with the use of SAE (lower part of Figure 3) for denoising of the HOG-dataset biometric data (Figure 2) obtained using FAP from the Caltech Faces database. HOG-dataset represented 19 faces of different people with 70 copies for each subject. The designed SAE consisted of 2 sparse autoencoders, that were trained and tuned with similar parameters and in such way as in previous experiment. The layer element dimensions of SAE scheme was chosen to be 511-100-50 ending with a Soft-Max classifier for 19 subjects (lower part of Figure 3). Unsupervised training and classification was carried out with transformed and masked HOG-samples [4] from Caltech Faces database with sizes in ratio 4:1. In these simulations SAE was used both for removing the "biometric noise" and for authentication of a person's face by its biometric HOG-vector with a length of 511 real numbers. The results obtained demonstrated that FRR (False Rejection Rate) takes the value 3.2% without application of error correcting codes as it was realized in Fuzzy Commitment scheme [4], where the application of BCH codes (511,58,91) and (511,28,111) allowed to achieve FRR at almost the same level of 3%.

Conclusion

In this work, an analysis of various types of autoencoders and the principles of their construction has been performed. A method for solving applied problems of processing image data with the use of extraction of all available features of images (video), and then reducing their dimensionality and removing noise with the use of autoencoders was proposed. The experiments based on of UvaNemo and Caltech datasets performed have shown the improvement in the accuracy of classification genuine smile from posed one by 10%, as well as a reduction in the complexity of biometric templates design. In addition, the hypothesis of biometric image data denoiseing from internal noise using SAE was confirmed. The prospects of research how to use the latent space of autoencoders has been determined.

Acknowledgment

This work was supported in part by the COST Action CA16101 "MULTI-modal Imaging of FOREnsic SciEnce Evidence tools for Forensic Science".

References

[1] M. Tschannen, O. Bachem, and M. Lucici, "Recent Advances in Autoencoder-Based Representation Learning," Proc. W. Bayesian DL. NeurIPS, 2018. arXiv preprint: https://arxiv.org/abs/1812.05069 (accessed 2019, Dec).

[2] P. Vincent, H. Larochelle, B. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," Proc. of the 25th ACM Int. conf. ML, 2008. pp. 1096–1103.

[3] B. A. Olshausen and D. J. Field, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1," Vision Research, Vol.37, 1997, pp.3311–3325.

[4] B. Assanovich and Yu. Veretilo, "Biometric database based on HOG structures and BCH codes," Proc. ITS, Minsk, 2017, pp. 286-287.

[5] T. Baltrusaitis, L.-P. Morency, and P. Robinson, "OpenFace: An open source facial behavior analysis toolkit," WACV, 2016, pp. 1-10.

[6] B. Assanovich, Yu. Veretilo, N. Bich, A. Pushkina, and V. Khilmanovich, "Recognition of Genuine Smile as a Factor of Happiness and its Application to Measure the Quality of Customer Retail Services," PRIP 2019, pp.84-89.

[7] H. Guo, X. H. Zhang, J. Liang, and W. J. Yan, "The Dynamic Features of Lip Corners in Genuine and Posed Smiles," J. Front.Psychol. Feb., 2018.

Автоэнкодеры для приложений шумоподавления и классификации

Ассанович Б. А.

Проанализировано несколько структур автоэнкодеров, используемых для эффективного кодирования данных с обучением без учителя и применяемых для решения задач классификации и удаления внутреннего шума из данных, используемых в задачах биометрического и эмоционального распознавания. Эксперименты по распознаванию типов улыбок и биометрической идентификации с использованием преобразованных характеристик данных из видео базы улыбок Smile UvA-NEMO и наборов данных Caltech Faces показали возможность повышения точности классификации на 10%.

Received 19.01.2020