International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 12 (2020) pp. 104-112


Received: 15 January, 2020; Accepted: 18 May, 2020; Published: 27 May, 2020

# Off-line Handwritten Arabic Text Recognition using Convolutional DL Networks

**Mohamed Elleuch[1] and Monji Kherallah[2]**

[1] National School of Computer Science (ENSI), University of Manouba, Tunisia
*elleuch.mohameds@gmail.com*

[2] Faculty of Sciences, University of Sfax, Tunisia
*monji.kherallah@fss.usf.tn*

***Abstract*: In recent decades many researchers worked on handwritten document analysis field and more specifically for Arabic Handwriting Script (AHS). Deep learning (DL) has revolutionized computer vision with several good examples, particularly the studies of the convolutional neural network on image classification. In this paper we investigate the benefit of deep convolution networks in textual image classification. Convolutional Deep Belief Networks (CDBN) is applied to learn automatically the finest discriminative features from textual image data consisting of AHS. This architecture is able to lay hold of the advantages of Deep Belief Network and Convolutional Neural Network. We subjoin Regularization methods to our CDBN model so that we can address the issue of over-fitting. We evaluated our proposed model on low and high-level dimension in Arabic textual (character /word) images using IFN/ENIT datasets with data augmentation. Experimental results show that our proposed CDBN architectures achieve better performance.**

***Keywords*:** Arabic Handwritten Script, Data augmentation, Deep convolution networks, CDBN, Regularization, over-fitting**.**

## I. Introduction and related works

Calculation techniques developed with Artificial Intelligence (AI) produce amazing results in solving many complex problems in different areas of our everyday life. Recent research has shown that deep learning (DL) techniques have made significant advances in solving problems such as computer vision, object recognition and natural language processing.

The techniques in relation to the information processing at present cognizes hectic progress in relationship with data processing. It has an increasing potential in the domain of the human-computer interaction. Furthermore, in recent years, human reading's machine simulation has been intensively subjected to many studies. The recognition of writing is part of the larger domain of pattern recognition. It aims at developing a system able to be the closest to the human ability of reading. Arabic-handwriting languages are lagging behind mainly because of their complexity and their cursive nature. Consequently, automatic recognition of handwritten script represents a burdening work to be fulfilled. Since the late $1960_s$, by dint of its broad applicability in several engineering technological areas, Arabic handwritten script (AHS) recognition has been positively seen as the subject of in-depth studies [1]. A lot of studies have been realized to recognize Arabic handwritten characters using unsupervised feature learning and hand-designed features [2] [3].

Improving suitable characteristics from the image describes a difficult and complex chore. It really requires not only a skilled but also an experienced specialist in the domain of feature extraction methods like: MFCC features in speech area, Gabor and HOG features in computer vision. The choice and goodness of these hand-designed features makes it possible to identify the efficiency of the frames utilized for classification and recognition like Multi-layer Perceptron (MLP), Hidden Markov Model (HMM), Support Vector Machine (SVM), etc. Many researchers have targeted the use of raw or untagged data in training developed handwriting systems, as they are the easiest way to handle large data.

Automatic feature extraction capability and high-level abstraction modeling in various signals namely text and image have made the deep learning algorithms prevalent in the world of Artificial intelligence research. Therefore, our first current study defy is to develop a richer automatic features extraction system than that obtained using heuristics signal processing based on the domain of knowledge. This approach is dependent on the notion of deeply learning a representation of the Arabic handwriting from the image signal. So as to carry out that, the uses of supervised and unsupervised learning methods have proved some potential. Learning such representations is likely to be applied to many handwriting recognition chores.

Recent research has shown that deep learning (DL) methods have made it possible to make decisive progress in solving tasks such as object recognition [4] [5], computer vision [6], speech recognition [7] [8] and Arabic handwriting recognition [9].

A deep architecture is made up of many layers of non-linear parameterized models (See Figure 1). These are parameters that are subject to learning. Each layer enables a higher level of representation than the last. We therefore hope to learn an abstract representation of data, in which the task to be solved is easier.

Elaborate by LeCun et al. [10], Convolutional Neural Network (CNN) is a specialist type of Neural Network (NN) that automatically learning favorable features at every layer of the architecture based on the given dataset, which can be a convolution layer, a pooling layer and a fully connected layer. Then Ranzato et al. [11] improved performance by using unsupervised pre-training on a CNN.

Another classifier which is employed extensively is Deep Belief Network (DBN) [12]. DBN is one of the most classical deep learning models, composed of several Restricted Boltzmann Machines (RBM) in cascade. This model learns representations of high-level features from unlabeled data that uses unsupervised learning algorithms.

In comparison to shallow learning, the pros of DL are that deep structures can be designed to learn internal representation and more abstract details of input data. However, the high number of parameters given can also lead to another problem: over-fitting. Thus, improving or developing novel effective regularization techniques is an unavoidable necessity. In recent years, various regularization techniques have been suggested as batch normalization, Dropout and Dropconnect.

Data augmentation is another way to minimize over-fitting on models, where we increase the amount of training data by using details only in training data. For this purpose, we use augmentation techniques applied in the IFN/ENIT database for: (1) reduce over-fitting and, (2) to best improve the classifier.

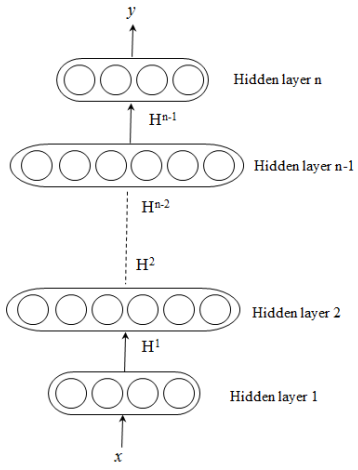



**Figure 1.** A Deep architecture topology

The participations of this paper are to leverage the DL approach to solve the problem of recognizing handwritten text in Arabic. To fulfill our target, we are studying the potential benefits of our suggested hybrid CDBN/SVM architecture [13]; this model handled CDBN as an automatic characteristic extractor and let SVM to be the output predictor (See Figure 2). On the other hand, to enhance the performance of CDBN/SVM model, regularization methods can contribute to the defense of over-fitting as Dropout and Dropconnect techniques.

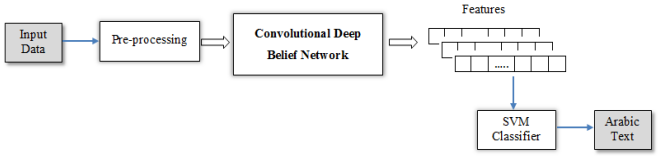



**Figure 2.** System based in CDBN Model

This paper is organized as follows: Section 2 gives an overview of the basic components of Convolutional Neural Network, Convolutional Deep Belief Network models and regularization techniques. Then, our target architectures are explored and discussed to recognize Arabic handwriting text. Section 3 describes experimental study and compares the results, and Section 4 discusses the results. In the final section, we conclude this work with some remarks and general perspectives.

## II. Deep models for handwritten recognition

In this section, the DBN model based on the RBM is firstly represented and after that, the CNN and CDBN models are reviewed. Just then, the effect of Dropout and Dropconnect techniques is analyzed in our CDBN architectures.

### *A. Restricted Boltzmann Machine (RBM)*

Deep Belief Networks is a deep generative model [12] involving several layers of RBM [14] consisting of a layer of visible units representing the input data and many layers of hidden units. The link between the two upper layers of DBN is not oriented, the other links are oriented, and there is no link for the neurons of the same layer (See Figure 3).

A Restricted Boltzmann Machine is a non-directed graphical model layer consisting of a visible units layer 'v', and a hidden units layer 'h', with a complete set of connections between them. The energy function as well as the joint distribution are written as follows:

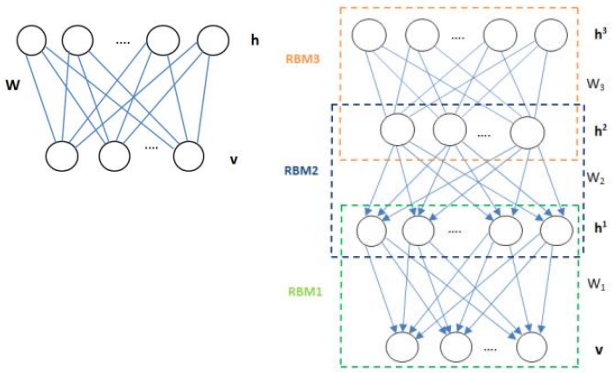
$$E(v,h) = -\sum_{i,j} v_i w_{ij} h_j - \sum_j b_j h_j - \sum_i c_i v_i \quad (1)$$

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)} \quad (2)$$

Where $w_{ij}$ constitute the synaptic weight between visible units $v_i$ and hidden units $h_j$, $b_j$ represent hidden unit biases, $c_i$ represent visible unit biases and $Z$ constitutes the partition function.

The energy function can be described otherwise as in Eq. (3) when the value of visible units is real:

$$E(v,h) = \frac{1}{2}\sum_i v_i^2 - \sum_{i,j} v_i w_{ij} h_j - \sum_j b_j h_j - \sum_i c_i v_i \quad (3)$$




(a) (b)

**Figure 3.** Representation of a RBM (a) and a DBN composed of three hidden layers (b)

### B. *Overview of CNN Classifier*

The first functional deep architectures to emerge were the Convolutional Neural Networks or CNNs [10] [15]. These models have proven their worth in the field of vision recognition.

In a CNN, there are two types of layers in alternating: a layer of convolution and a layer of sub-sampling (pooling). The layer of convolution consists of a number of filters having a limited field of perception. These filters work as detectors of features (See Figure 4). They are defined by weights that are adjustable during learning. Performing a convolution of these filters on the input for the activation of the layer. One can see the convolution filters as a sharing of weight between units of the same filter applied to different locations. Activations are then sent to the layer of sub-sampling. Thereof, also formed of units having a limited field of perception, reduces the size of the previous layer by a function of sub-sampling (typically the average or maximum activations in the perceptual field). The next layer then performs a convolution on the outputs of the layer of sub-sampling, and so on. After the last layer of sub-sampling, a layer of output neurons is added, allowing us to perform supervised training by retro-propagating gradient.

The alternation between convolution and sub-sampling provides a pyramid structure to CNN. Consequently, the upper layers represent feature increasingly global of input because their perceptual field, although low dimensionality corresponds to a larger portion of the input. In addition, weight sharing allows to limit the number of model parameters facilitating learning.

The convolution operation is defined as:

$$x_j^l = f\left(\sum_i y_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{4}$$

Where

- $x_j^l$ represents the *j-th* convolution result of the *l-th* layer,
- $y_i^{l-1}$ is the *i-th* output map of its previous layer,
- '*' denoted the convolution operation,
- $k_{ij}^l$ represents the convolution kernel and $b_j^l$ is the bias of the *j-th* output layer

Non linearity has been performed in CNN by activation functions and pooling layers. We exploited the activation function ReLU (Rectified Linear Units):

$$f(x) = \max(0, x) \tag{5}$$

The FCL takes the function written as follows:

$$y_j^l = \max\left(0, \sum_i y_i^{l-1}.w_{i,j}^l + b_j^l\right) \tag{6}$$

Where $y_j^l$ represents the *j-th* node in the *l-th* layer, $w_{i,j}^l$ represents the weights between $y_j^l$ and $y_i^{l-1}$, and $b_j^l$ is the bias.

The last step of foretelling a distribution $P(y_i)$ is to handle a softmax over the outputs $z_i$ (shapes of characters):

$$y_i = \frac{\exp(z_i)}{\sum_k \exp(z_k)} \tag{7}$$

$$z_i = \sum_j y_j^{l-1}.w_{i,j} + b_i \tag{8}$$

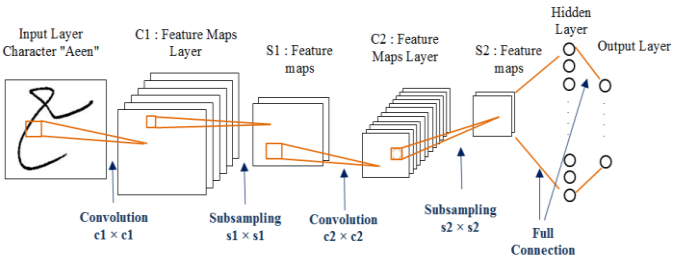



**Figure 4.** A typical CNN architecture

### C. *Convolutional Restricted Boltzmann Machine (CRBM)*

The construction of hierarchical features structures is a challenge and the Convolutional DBN is one of the famous features extractor often used in the last decade in the field of pattern recognition. In this subsection, we thoroughly clarify the basic notion of this approach.

As a hierarchical generative model [16], the Convolutional Deep Belief Network reinforces the efficiency of bottom-up and top-down probabilistic inference. Similar to the DBN standard, this model made up of several layers of max-pooling CRBMs stack on top of each other, and the training was carried out by the greedy layer-wise algorithm [17] [12]. Building a convolutional DBN, the algorithm learns high-level features, such as groups of the strokes and part of the object. In our experiments, we trained Convolutional DBN with a couple of CRBM layers and, for the inference we utilized the feed-forward approximation [13].

Convolutional Restricted Boltzmann Machine is the foundation of Convolutional DBN. We train the Convolutional DBN model by learning a CRBMs stack in which a CRBM output matches the next CRBM input on the stack. Figure **5** clarifies the architecture of CRBM made up of two layers: a visible layer V and a hidden layer H, both joined by sets by sets of local and common parameters. The visible units are assessed in binary or real value and the hidden units in binary value.

Supposing that the visible input layer consists of L images in where each consists of $N_V \times N_V$ real units (image in intensity pixels). The hidden units H are divided into K maps or groups, each map being an $N_H \times N_H$ array of a set of binary units and associated with a $N_W \times N_W$ convolutional filter (where $N_W \triangleq N_V - N_H + 1$). We share filter weights among all hidden unit locations in the same group. There is a shared bias $b_k$ for each map and a shared bias $c$ for visible units alike.

A new process for the Convolutional DBN structure called "probabilistic max-pooling" was improved by Lee et al. [16]. This probabilistically decreases the representation of the detection layers. Decreasing the representation with max-pooling allows representations of the upper layer to never change to local translations of input data, reduces the computational load [18] and is useful for vision recognition issues [19]. By using

visible inputs with real values, the probabilistic max-pooling CRBM is fixed by the following equation:

$$E(v,h)=\frac{1}{2}\sum_{i,j=1}^{N_V}v_{i,j}^2-\sum_{k=1}^{K}\sum_{i,j=1}^{N_H}\sum_{r,s=1}^{N_W}h_{i,j}^k w_{r,s}^k v_{i+r-1,j+s-1} \quad (9)$$
$$-\sum_{k=1}^{K}b_k\sum_{i,j=1}^{N_H}h_{i,j}^k-c\sum_{i,j=1}^{N_V}v_{i,j}$$

From equation (9), the conditional and joint distributions are computed as:

$$P(v,h)=\frac{1}{Z}\exp(-E(v,h)) \quad (10)$$

$$P(v_{i,j}=1|h)=N((\sum_k W^k *_f h^k)_{i,j}+c,1) \quad (11)$$

$$P(h_{i,j}^k=1|v)=\frac{\exp\left(I(h_{i,j}^k)\right)}{1+\sum_{(i',j')\in\beta_\alpha}\exp\left(I\left(h_{i',j'}^k\right)\right)} \quad (12)$$

Where

- $I(h_{i,j}^k)\triangleq b_k+(\tilde{W}^k *_v v)_{i,j}$ represent the data hidden units in map $k$ got from visible layer $V$, $\tilde{W}$ fixed as the matrix filter $W$ overturned at once left-right directions and up-down side, $*_v$ represent a valid convolution,
- $*_f$ represent a complete convolution and $N(.)$ represent a normal distribution,
- $\beta_\alpha$ refers to a $C \times C$ block denoted α where is pooled or connected to a binary node $P_\alpha^k$ in the pooling layer. The pooling node $P_\alpha^k$ is fixed as $P_\alpha^k\triangleq\sum_{(i',j')\in\beta_\alpha}h_{i,j}^k$ and the conditional probability is determined by:

$$P(p_\alpha^k=1|v)=\frac{\sum_{(i',j')\in\beta_\alpha}\exp\left(I(h_{i',j'}^k)\right)}{1+\sum_{(i',j')\in\beta_\alpha}\exp\left(I\left(h_{i',j'}^k\right)\right)} \quad (13)$$

Exploiting the operators defined before, the energy function in equation (9) can be recalculated as follows:

$$E(v,h)=\frac{1}{2}\sum_{i,j=1}^{N_V}v_{i,j}^2-c\sum_{i,j=1}^{N_V}v_{i,j}-\sum_{k=1}^{K}\sum_{i,j}\left(h_{i,j}^k\left((\tilde{w}^k *_v)_{i,j}\right)+b_k\right) \quad (14)$$

In the same way as RBM, training CRBM is performed by exploiting the Contrastive Divergence (CD) procedure [12], which is an approximation of the maximum likelihood estimate. Moreover, CD allows us to determine an approximate gradient in an efficient way [20]. The Learning algorithms and inference are based on the block Gibbs sampling method. After forming a max-pooling CRBM we can use it to calculate the posterior of the hidden units based on the input data (pooling). These hidden unit "activations" can be exploited as input to further train the next layer CRBM. Algorithm 1 describes the training algorithm steps for max-pooling CRBM [16].

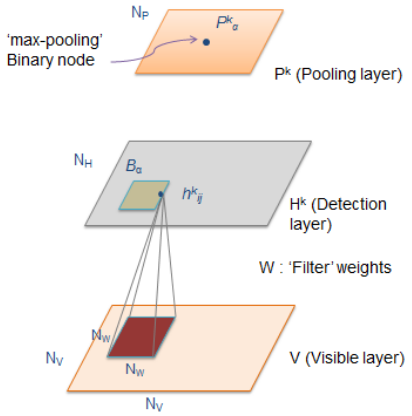



**Figure 5.** Representation of a probabilistic max-pooling CRBM. $N_V$ and $N_H$ refer to the dimension of visible and hidden layer, and $N_W$ to the dimension of convolution filter

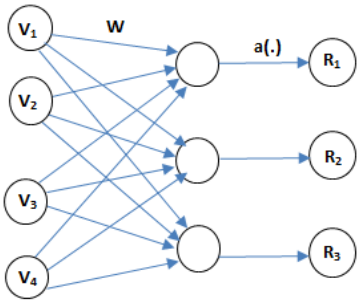
**Algorithm 1: Training algorithm for Convolutional RBM**

**Input:** Sequence of $L$ image examples; $<(N_v \times N_v)_1, \ldots, (N_v \times N_v)_L>$ matrix $V$
Initialize of CRBM parameters: learning rate, filter size, maps, bias $\{b_k, c\}$ and filters $W_k$
**Output:** New filters $W_k$
Repeat the steps 1 to 4 {to $L$ training images} until convergence
(1) $V_0 \leftarrow V$
Compute the posterior $Q_0$ using equation 7 and 8
(2) Sample $H_0$ from $Q_0$
(3) For $n \leftarrow 1$ to $K$ do
Sample $V_n$ from $P(V|H_{n-1})$ using equation 6
Compute the posterior $Q_n$ using equation 7 and 8
Sample $H_n$ from $Q_n$
End for
(4) Update weights and biases with contrastive divergence and sparsity regularization

### D. *Regularization methods*

The utilization of Deep Networks models for cursive handwriting recognition has made significant progress over the past decade. Nevertheless, for these architectures to be used effectively, a wide amount of data needs to be collected.
Consequently, over-fitting is a serious problem in such networks due to the large number of parameters that will be carried out gradually as the network increases and gets deeper. To overcome this problem, many regularization and data augmentation procedures have been ameliorated [21] [22] [23].
In this sub-section, two regularization techniques will be shortly introduced that may affect the training performance. Dropout and Dropconnect are both methods for preventing over-fitting in a neural network.
To practice Dropout, a subset of units are haphazardly selected and set their output to zero without paying attention to the input. This efficiently removes these units from the model. A Varied subset of units is selected randomly each time we present an example of training. Figure 6 demonstrate the variance between No-Drop and Dropout networks.

**(a)**



**(b)**

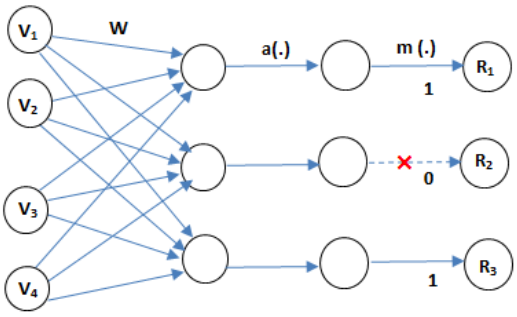



**Figure 6.** (a) No-Drop network (b) Dropout network. Where, m is weight mask, a(.) is the activation function, W is layer weights and V is layer inputs

Dropconnect works similarly, except that we disable individual weights (i.e., set them to zero), instead of nodes, so a node can remain partially active. Furthermore, Dropconnect is a generalization of Dropout because it produces even more possible models, since there are almost always more connections than units (See Figure **7**).

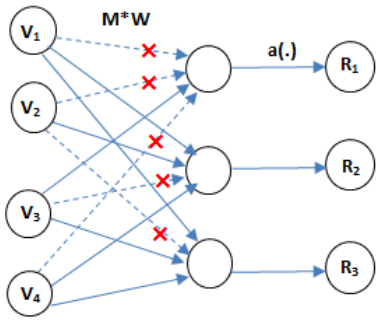



**Figure 7.** DropConnect Network

*E. Model settings*

To extend our study [13] [24] so that we can discover the power of the deep convolutional neural networks classifier done on the problem of Arabic handwritten script recognition, we point out in this work an itemized study of CDBN with Dropout/Dropconnect techniques. In this subsection, we identify the tuning parameters of the chosen CDBN structure.

As noted above, our CDBN architecture is composed of two layers of CRBM (See figure 8). The efficiency of this architecture during IFN/ENIT's handwritten text recognition task was evaluated.

The description of the CDBN architecture exploited in the experiments conducted in the IFN/ENIT database is given as follows: $1 \times 300 \times 100 - 12W24G - MP2 - 10W40G - MP2$. This architecture corresponds to a network with dimension input images $300 \times 100$, the initial layer consisting of 24 groups of $12 \times 12$ pixel filters and the pooling ratio C for each layer is 2. The second layer includes 40 maps, each$10 \times 10$. We define a sparseness parameter of 0.03. The initial layer bases learned strokes consisting of the characters, as for the second layer bases learned characters parts by the groups of strokes. By integrating the activations of the first and second layers, we constructed feature vectors; Support vector machines are used to rank these features.

On the other hand, for handwritten characters recognition task we train the subsequent CDBN architecture configuration: $1 \times 50 \times 50 - 11W40G - MP2 - 11W40G - MP2$ stands for a network with input images of dimension $50 \times 50$, the initial and the second layer are made up of 40 groups of $11 \times 11$ pixel filters and the pooling ratio *C* for every layer is 2. Every group in the hidden layer is linked to a pooling unit. We employed 0.05 as a low density target for the initial and second layer to set the number of activation weight at one time.

In order to regularize and make the most effective use of these architectures, units or weights have been removed. Dropout was used only at the input layer with a probability of 20% and at each hidden layer at a probability of 50 %, while Dropconnect was only applied at the input layer with a probability of 20%.

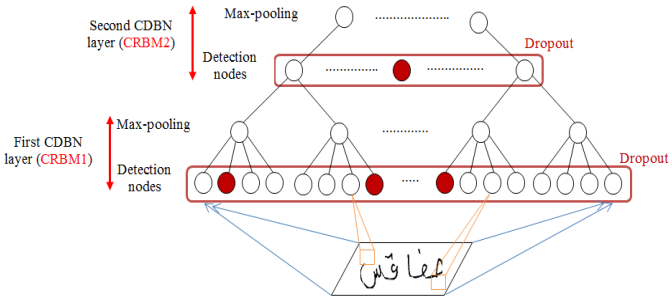



**Figure 8.** Representation of the suggested CDBN structure with dropout

*F. Support Vector Machines (SVM)*

Invented by Vapnik [25] and Cortes [26] support vector machines, have turned into being a popular approach used in various domains. It is regarded as be the state-of-the-art tool for resolving linear and non-linear classification problems (See Figure 9). SVM provides a high generalization performance. Moreover, it uses the Structural Risk Minimization principle and it tries to keep away the over-fitting problem by obtaining the decision hyper-plane which is most favorable to the maximum margin between classes.

The most favorable hyper-plane is acquired by solving a quadratic programming problem subject to regularization parameters. The linear SVM algorithm has been prolonged to transaction with non-linear classification problems. To elaborate with non-linear decision boundaries, the best way is to transform $x_i$ to a higher dimension space using a transformation function Φ, so as to, in this new space, the samples are plausible to be linearly separable. By using Kernels, SVM work out these difficulties. Many non-linear kernel functions have been suggested. The kernel function is termed as:

$$K(x_i, x_j) = \Phi(x_i) \times \Phi(x_j) \qquad (15)$$

As examples, linear, Radial Basis Function (RBF), Sigmoid and Polynomial kernel types are defined as;

- The linear kernel: $K(x_i, x_j) = x_i \times x_j$
- The polynomial kernel: $K(x_i, x_j) = [(x_i \times x_j) + 1]^d$
- The Sigmoid kernel: $K(x_i, x_j) = tanh(\beta_0 x_i x_j + \beta_1)$
- RBF kernel (Radial Basis Function): $K(x_i, x_j) = exp(-\gamma || x_i - x_j ||^2)$

$\gamma = 1 / \sigma^2$

With $d$, $\beta_0$, $\beta_1$, and $\gamma$ are parameters that will be determinate empirically.

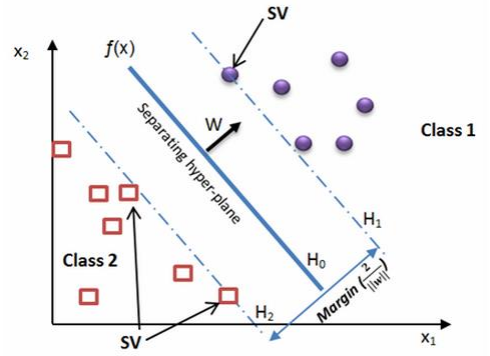



**Figure 9:** Principle of SVM; two-class hyper-plane example, SV are Support Vectors and $H_0$ is the optimal hyper-plane, defined according to that, in order to maximizes the margin i.e. the distance to the classes Class 1 and Class 2

## III. Experiments with proposed model

This section illustrates a test to evaluate the suggested approach performance on the IFN/ENIT benchmark database [27]. In our experiments, each IFN/ENIT dataset image was normalized to the same input dimension with 300×100 pixels for the visible layer, whilst in the generated characters database the images are resized to 50×50 pixels. These textual images are at the gray level and resizing is not necessarily square.

The problem with the IFN / ENIT database is that each class is made up of small datasets. Therefore, the models trained with them do not generalize well the data from the validation and test set. Consequently, these models suffer from the problem of over-adjustment. To overcome this issue, we resort to data augmentation.

- *Pre-processing:* This phase is to generate standardized text image and uniform.
- *Data augmentation:* DL methods generally need a great amount of training data in order to obtain a good learning of all the parameters associated at each layers of a deep network. Therefore, augmentation must be achieved by transforming the original data. The main techniques fall under the category of data warping [28] [29].
- *Parameters setting:* For configuration, it is a must to identify the number and size of filters, sparsity of the hidden units and max-pooling region size in each layer of the CDBN model. Referring to the size of the images used (low and high-dimensional data), we specify a hyper-parameters setting for the configuration of the CDBN structure. So, to get the most out use of this architecture, two regularization methods have been put into practice separately for the CDBN structure called Dropout and DropConnect.

### A. Dataset description and experimental setting

To measure the effectiveness of our system proposed for low and high-level dimension of data, the IFN/ENIT database [27] is employed. Indeed, the IFN/ENIT database comprises 32492 handwritten Arabic words developed with contributions from 1000 volunteers, making a total of about 138060 parts of Arabic words (PAWs) and about 257366 characters (See Table 1). The words written are 946 Tunisian town and village names with the postal code of each. The data treatment is made up of offline handwritten Arabic words. Datasets 'a-b-c' are used for training phase whereas the test set was selected from set 'd'. Figure 10 illustrates samples of village name, written by 5 different writers. On the other hand, words are segmented into letters (See Figure 11) from set (a) to (e). We have kept 3.360 images as train data and 1.120 images as test data. These images include 56 shapes of characters. Details of the class for each shape are presented in Table 2.

| Set | Number of words | Number of characters | Number of PAWs |
|---|---|---|---|
| a | 6537 | 51984 | 28298 |
| b | 6710 | 53862 | 29220 |
| c | 6477 | 52155 | 28391 |
| d | 6735 | 54166 | 29511 |
| e | 6033 | 45169 | 22640 |
| *Total* | *32492* | *257366* | *138060* |

*Table 1.* Different IFN/ENIT Datasets.

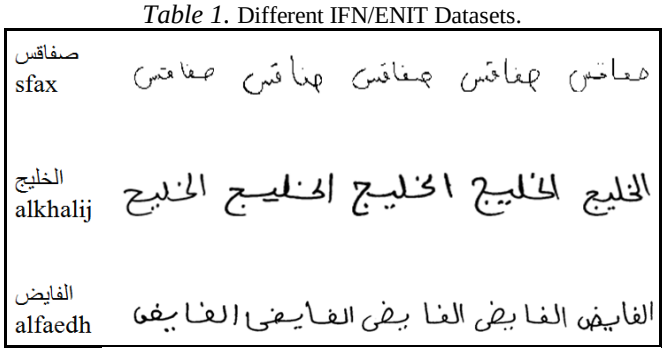



**Figure 10.** Examples of Arabic words from the IFN / ENIT data set

| Arabic Script | Shape | class | Arabic Script | Shape | class |
|---|---|---|---|---|---|
| Aeen (ع) | ع | 1 | Laam (ل) | ل | 29 |
| | عـ | 2 | | لـ | 30 |
| | ـعـ | 3 | | ـل | 31 |
| | ـع | 4 | | ـلـ | 32 |
| Alif (ا) | ا | 5 | Lam_Alif (لا) | لا | 33 |
| | ـا | 6 | | ـلا | 34 |
| Baa (ب) | بـ | 7 | Meem (م) | مـ | 35 |
| | ـبـ | 8 | | ـمـ | 36 |
| | ـب | 9 | | ـم | 37 |

| | | | | | |
|---|---|---|---|---|---|
| | | 10 | | | 38 |
| Daal (د) | | 11 | Noon (ن) | | 39 |
| | | 12 | | | 40 |
| Faa (ف) | | 13 | Raa (ر) | | 41 |
| | | 14 | | | 42 |
| | | 15 | Saad (ص) | | 43 |
| | | 16 | | | 44 |
| Haa (ه) | | 17 | | | 45 |
| | | 18 | | | 46 |
| | | 19 | Seen (س) | | 47 |
| | | 20 | | | 48 |
| | | 21 | | | 49 |
| Hamza (ء) | | 22 | | | 50 |
| Jeem (ج) | | 23 | Taa (ط) | | 51 |
| | | 24 | | | 52 |
| | | 25 | Waao (و) | | 53 |
| | | 26 | | | 54 |
| Kaaf (ك) | | 27 | Yaa (ى) | | 55 |
| | | 28 | | | 56 |

*Table 2.* Class for each shape of an Arabic script.

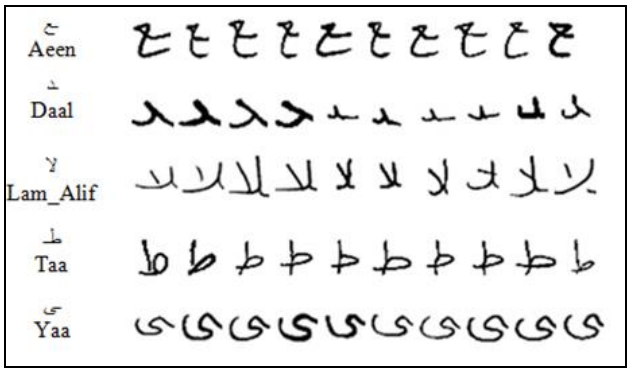



**Figure 11.** Samples of Arabic letters generated from the IFN/ENIT database.

*B. Experimental results and comparison*

Table 3 makes a comparison between our approach outcomes with those already published results. We noted that the work of our CDBN structure yielded promising results, with a Word Error Rate (WER) of about 8.45 % if compared to Maalej and kherallah's works [30], after applying Dropout. On the other hand, with Dropconnect we got an error rate of 13.85 %.

In addition, the rate obtained is contrasted to our previous work. These experiments clearly prove that the outcome in [13] reaches 16.3% using the CDBN structure without Dropout, which is not excellently contrasted to the classic method [31] [32]. It is thanks to the CDBN model that is able to be over-completed. On an experimental basis, a model that is too complete or too adjusted may be prone to learn insignificant solutions, such as pixel detectors. In our current work to find a suitable solution to this problem, we utilize two regularization techniques, namely Dropout and Dropconnect for CDBN. As a result, the acquired outcomes prove an improvement rate of approximately 7.85 % with Dropout and 2.45 % with Dropconnect.

A second experimental study was established on Arabic letters. The Character Error Rate (CER) achieved 5.26% (59 erroneous characters) by CDBN with Dropconnect, whereas adding Dropout reduces the error to 4.46% (50 erroneous characters). So, our CDBN model with dropout or Dropconnect outperforms CNN based SVM classifier [33].

| **Authors** | **Approach** | **WER (train/test sets)** | **CER** |
|---|---|---|---|
| **Present work** | CDBN with Dropout | **8.45%** (a-c/d) | **4.46** % |
| | CDBN with Dropconnect | 13.85 % (a-c/d) | 5.26 % |
| Elleuch and Kherallah, 2019 [34] | CDBN with Dropout | 7.1% (a-d/e) | -- |
| Elleuch et al., 2016 [33] | CNN/SVM with Dropout | -- | 7.05% |
| Maalej and Kherallah, 2016 [30] | RNN (MDLSTM with dropout) | 11.62 % (a-c/d) | -- |
| Elleuch et al., 2015 [13] | CDBN (without dropout) | 16.3 % (a-c/d) | -- |
| AlKhateeb et al., 2011 [31] | HMM | 13.27 % (a-d/e) | -- |
| Amrouch et al., 2018 [32] | CNN-HMM | 10.77 % (a-c/d) | -- |

*Table 3.* Performance comparisons approaches utilizing the IFN / ENIT database.

In general, it is evident that the proposed deep learning architecture, Convolutional Deep Belief Network with Dropout [34], provides satisfactory performance, specially when compared over others approaches such as the Recurrent Neural Networks (RNN) and the HMM method applied to the IFN/ENIT database. We also compare the results of models trained with and without augmented data. Clearly results using data augmentation are much better than the results without using augmentation.

## IV. Discussion

As mentioned above, our suggestion describes a deep learning approach for AHS recognition, in particular the Convolutional Neural Network and Convolutional Deep Belief Network. To validate the efficiency of the proposed framework, we presented experimental outcomes utilizing Arabic words/characters handwritten databases with augmented data; IFN/ENIT database.

We are able to observe that our CDBN architecture with Dropconnect has reached a promising error rate of 13.85 % when used with large dimension data. In addition, we have rebuilt our proposed CDBN configuration with Dropout. The performance is then raised to reach a WER of 8.45 %, which corresponds to a gain of 5.4%.

Furthermore, we have reconstructed our suggested CDBN structure configuration to enable it to process with low dimension data as characters and assess on the IFN/ENIT database with augmented images. The performance is then increased to reach the rate of 94.74% using Dropconnect (CER of 5.26%). While applying Dropout the obtained rate achieved a promising accuracy rate of 95.54% (CER of 4.46%).

The results obtained, regardless of their size, are sufficiently important compared to scientific researches using other classification methods, in particular those they obtained with raw data without feature extraction phase (See figure 12). This participation portrays an interesting challenge in the field of computer vision and pattern recognition, as it will be a real incentive to motivate the use of deep machine learning with Big Data analysis.

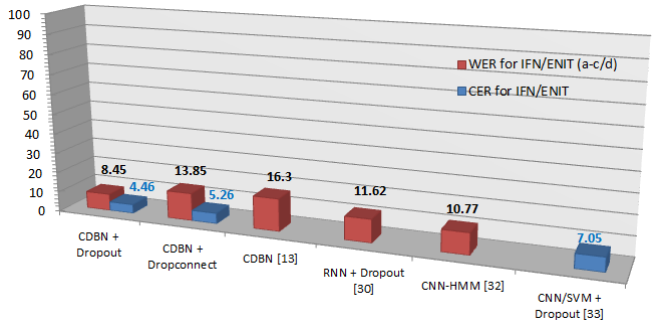



**Figure 12.** CER / WER comparison using IFN/ENIT Database

## V. Conclusion

With the development of DL technique, deep hierarchical neural network has drawn great attentions for handwriting recognition. In this article, we first introduced a baseline of the DL approach to AHS recognition, primarily the Convolutional Deep Belief Network. Our aim was to leverage the energy of these Deep Networks that can process large dimensions input, permitting the usage of raw data inputs rather than extracting a feature vector and studying the complex decision boundary between classes. Secondly, we studied the performance of two regularization methods applied separately in the CDBN structure to recognize Arabic words/characters using IFN/ENIT Database with augmented images. As we can observe, Dropout is a very efficient regularization technique compared to Dropconnect and the unregulated basic method. Still, it can be concluded that data augmentation makes it possible to reduce the over-fitting of models.

In future works, we have to practice other regularization techniques such as Maxout for training and to utilize reinforcement learning algorithms as deep Q-networks to increase performance and further improve the accuracy rate of recognition. We plan also to investigate other models such as VGGNet and RestNet.

## References

[1] Mota, R., and Scott, D., 2014, "Education for innovation and independent learning,"

[2] Porwal, U., Shi, Z., and Setlur, S., 2013, "Machine learning in handwritten Arabic text recognition," In *Handbook of Statistics* (Vol. 31, pp. 443-469). Elsevier.

[3] Elleuch, M., Hani, A., and Kherallah, M., 2017, "Arabic handwritten script recognition system based on HOG and gabor features," *Int. Arab J. Inf. Technol.*, 14(4A), 639-646.

[4] Boureau, Y. L., and Cun, Y. L., 2008, "Sparse feature learning for deep belief networks," In *Advances in neural information processing systems* (pp. 1185-1192).

[5] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. A., 2010, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research*, *11*(Dec), 3371-3408.

[6] Huang, G. B., Zhou, H., Ding, X., and Zhang, R., 2011, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(2), 513-529.

[7] Mohamed, A. R., Dahl, G. E., and Hinton, G., 2011, "Acoustic modeling using deep belief networks," *IEEE transactions on audio, speech, and language processing*, *20*(1), 14-22.

[8] Dahl, G., Mohamed, A. R., and Hinton, G. E., 2010, "Phone recognition with the mean-covariance restricted Boltzmann machine," In *Advances in neural information processing systems*(pp. 469-477).

[9] Al-Ayyoub, M., Nuseir, A., Alsmearat, K., Jararweh, Y., and Gupta, B., 2018, "Deep learning for Arabic NLP: A survey," *Journal of computational science*, 26, 522-531.

[10] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., 1998, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, *86*(11), 2278-2324.

[11] Marc'Aurelio Ranzato, F. J. H., Boureau, Y. L., and LeCun, Y., 2007, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," In *Proc. Computer Vision and Pattern Recognition Conference (CVPR'07). IEEE Press* (Vol. 127).

[12] Hinton, G. E., Osindero, S., and Teh, Y. W., 2006, "A fast learning algorithm for deep belief nets," *Neural computation*, *18*(7), 1527-1554.

[13] Elleuch, M., Tagougui, N., and Kherallah, M., 2015, "Deep learning for feature extraction of Arabic handwritten script," In *International Conference on Computer Analysis of Images and Patterns* (pp. 371-382). Springer, Cham.

[14] Mohamed, A. R., Sainath, T. N., Dahl, G. E., Ramabhadran, B., Hinton, G. E., and Picheny, M. A., 2011, "Deep Belief Networks using discriminative features for phone recognition," In *ICASSP* (pp. 5060-5063).

[15] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J., 1989, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, *37*(3), 328-339.

[16] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y., 2011, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," *Communications of the ACM*, *54*(10), 95-103.

[17] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H., 2007, "Greedy layer-wise training of deep networks," *In Advances in neural information processing systems* (pp. 153-160).

[18] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y., 2009, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," In *Proceedings of the 26th annual international conference on machine learning* (pp. 609-616). ACM.

[19] Jarrett, K., Kavukcuoglu, K., and LeCun, Y., 2009, "What is the best multi-stage architecture for object recognition?," In *2009 IEEE 12th international conference on computer vision* (pp. 2146-2153). IEEE.

[20] Carreira-Perpinan, M. A., and Hinton, G. E., 2005, "On contrastive divergence learning," In *Aistats* (Vol. 10, pp. 33-40).

[21] Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012, "Imagenet classification with deep convolutional neural networks," In *Advances in neural information processing systems* (pp. 1097-1105).

[22] Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., and Fergus, R., 2013, "Regularization of neural networks using dropconnect," In *International conference on machine learning* (pp. 1058-1066).

[23] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R., 2012, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*.

[24] Elleuch, M., and Kherallah, M., 2019, "Convolutional Deep Learning Network for Handwritten Arabic Script Recognition," In *International Conference on Hybrid Intelligent Systems* (In press). Springer.

[25] Vapnik V., "*Statistical Learn Theory*," John Wiley, New York, 1998.

[26] Cortes, C., and Vapnik, V., 1995, "Support-vector networks," *Machine learning*, *20*(3), 273-297.

[27] Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., and Amiri, H., 2002, "IFN/ENIT database of handwritten Arabic words," *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, (pp. 127-136).

[28] Simard, P. Y., Steinkraus, D., and Platt, J. C., 2003, "Best practices for convolutional neural networks applied to visual document analysis," In *Icdar* (Vol. 3, No. 2003).

[29] Wigington, C., Stewart, S., Davis, B., Barrett, B., Price, B., and Cohen, S., 2017, "Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network," In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 639-645). IEEE.

[30] Maalej, R., and Kherallah, M., 2016, "Improving MDLSTM for offline Arabic handwriting recognition using dropout at different positions," In *International conference on artificial neural networks* (pp. 431-438). Springer, Cham.

[31] AlKhateeb, J. H., Ren, J., Jiang, J., and Al-Muhtaseb, H., 2011, "Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking," *Pattern Recognition Letters*, *32*(8), 1081-1088.

[32] Amrouch, M., Rabi, M., and Es-Saady, Y., 2018, "Convolutional feature learning and CNN based HMM for Arabic handwriting recognition," In International conference on image and signal processing (pp. 265-274). Springer, Cham.

[33] Elleuch, M., Maalej, R., and Kherallah, M., 2016, "A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition," *Procedia Computer Science*, *80*, 1712-1723.

[34] Elleuch, M., and Kherallah, M., 2019, "Boosting of Deep Convolutional Architectures for Arabic Handwriting Recognition," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, *10*(4), 26-45.

## Author Biographies

**Mohamed Elleuch** received the B.S. degree in computer science from The Faculty of Economics and Management of Sfax-Tunisia (FSEGS) and completed his master's degree in New Technologies of Dedicated Computer Systems from the National School of Engineering of Sfax (ENIS), Tunisia. He obtained the Ph.D. degree in Computer Science at the National School of Computer Science (ENSI), University of Manouba-Tunisia, in 2016. His main research concerns pattern recognition, handwriting recognition, neural networks, computer vision, image and signal processing.

**Monji Kherallah** graduated in Electrical Engineering 1989, obtained a Ph.D. in Electrical Engineering in 2008. He is now a professor in Electrical & Computer Engineering at the University of Sfax. His research interest includes applications of intelligent methods to pattern recognition and industrial processes. He focuses his research on handwritten documents analysis and recognition, handwritten Arabic recognition, biometrics, pattern recognition and image processing. He is member of the editorial board of "Pattern Recognition Letters." He was a member of the organization committee of the International Conference on Machine Intelligence ACIDCA-ICMI'2005.