Role of Deep Neural Features Vs Hand Crafted Features for Hand Written Digit Recognition

Jyostna Devi Bodapati, B Suvarna, Veeranjaneyulu N

ABSTRACT--- ***Handwritten digit recognition can be considered as a subtask of hand written character recognition, a broad area where a given character is recognized automatically by a machine. The major challenges of hand written character recognition are: writing style and size of characters varies from person to person. With the advances in machine learning algorithms the success of handwritten character recognition is improved. In this task we have considered hand written digit recognition, as there are plenty of real-time applications like amount identification on Bank cheques, recognizing zip codes on postal letters to mention few. Recent literature shows that performance of Convolution Neural Network (CNN) is promising on images. We have used neural network based models for hand written digit classi-fication. Initially the model is trained on MNIST dataset. In this work we have tried to identify the effect of different types of features on the performance of the model.***

Keywords: Deep learning, CNN, hand crafted features, hand written character recognition, pooling, convolution, dropout

I. INTRODUCTION

Pattern recognition is the process of identifying similar patterns in the given data. This area received increased demand from researchers in the field as there are plenty of real-time applications. In the area of pattern recognition there are many subfields like: Classification, Clustering, regression and dimensionality reduction. These are all highly demanding tasks in many of the real-time applications [1]. These tasks can be classified as either supervised or unsupervised depending on whether any supervised information is provided while training or not. As supervised information is being used while training, classification and regression tasks fall under the category of supervised learning. In case of clustering and dimensionality reduction tasks, supervised information is not available during training the input data so these tasks fall under the category of unsupervised learning. Regression is a function approximation task in which a continuous value is to be assigned to a given data point. Actually Classification can be considered as special type of regression. In regression the output of the test data is continuous on the other hand the output of classification task is discrete. To accomplish classification task, models like simple K-nearest neighbourhood(KNN), Gaussian mixture models (GMM)-based Baye's classification, artificial neural networks (ANN)- based Multilayer feed forward neural networks (MLFFNN), support vector machine (SVM) based classification are the most popular among other classification models [2].

Importance of features in classification task:

Data acquisition is the first and foremost job in any pattern recognition task. Once the data is collected then apply some pre-processing and cleaning techniques on the data and extract features. If required dimension reduction techniques are to be applied to reduce the number of features on the data. Then divide the data into train, validation and test data sets. On the train data, apply the intended model and tune the parameters of the model till we get desired accuracy on the validation data. In the final stage, to understand performance of the trained model, it has to be applied on the test data.

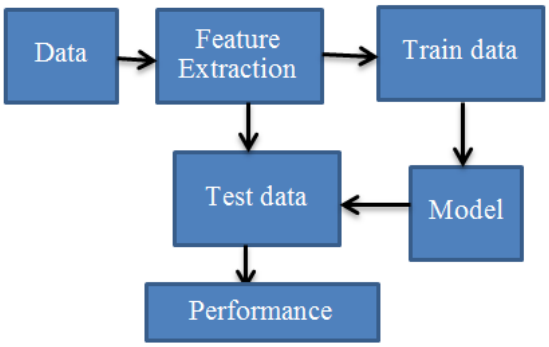
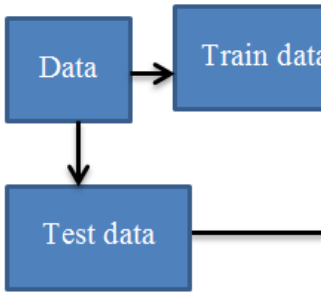


Figure 1: Flow diagram of Classification task in shallow model



Revised Manuscript Received on February 11 , 2019.

Jyostna Devi Bodapati Asssistant Professor, CSE Dept, Vignan's University , Guntur, Andhra Pradesh, India.

B Suvarna Asssistant Professor, CSE Dept, Vignan's University, Guntur, Andhra Pradesh, India. (E-mail: jyostna.bodapati82@gmail.com)

Veeranjaneyulu N Professor, IT Dept, Vignan's University, Guntur, Andhra Pradesh, India.

Figure 2: Flow diagram of deep learning based classification

In shallow models different algorithms are used to extract features and classification. The performance of these classifiers depends on the quality of features that were fed into the model. Features are the set of attributes of the data elements based on which the data elements can be classified into one of the available classes. Most widely used features for the image classification task are: Histogram of Gradients (**HOG),** gist and Scale invariant feature transform (sift) features [4]. At a high level, all these types of features compute image gradients (orientations and magnitude), break the image region into spatial bins, and finally histogram the gradient magnitudes according to their orientations and spatial location. The major difference lies in how these things are used. **HOG** is typically used in a **sliding window** fashion in object detection systems. **GIST** usually extracts features from the entire image and a single feature vector is used to represent an image. Gist features represent the whole image and a single feature vector is sufficient to represent the entire image hence called the global feature representation of the image. On the other hand, SIFT features are extracted from a local region. That is to represent an entire image, a set of local feature vectors are used. The number of local features to represent different data points need not be the same and could be varying. Hence these features for varying length pattern representation. The performance of these shallow classifiers is subject to how good the features are extracted from the data.

Recent literature claims that impressive gain in performance is possible using artificial neural networks for various classification tasks. As per the theory, single perceptron can classify the data that is linearly separable, where the classes are separable by a linear boundary. To solve more complex problems, where the boundary is a complex non-linear boundary then the single neural is not sufficient. Multi-layer feed forward neural network (MLFFNN) can be used to derive such complex non-linear boundaries required to separate complex data. MLFFNN also requires hand-crafted features as the input and requires multiple layers in the architecture to create a complex decision boundary. The network that is too deep in terms of the number of layers is called the deep neural network (DNN) and almost ruined all the shallow networks that were popular earlier. DNN suffered from vanishing gradients problem till the introduction of ReLU (Rectified linear unit) activation function. With the introduction of ReLU, performance of the DNN is increased with the number of layers as long as sufficient data is provided to the network. Still performance of DNN is dictated by the quality of features.

In 1998 Lecun proposed Convolution neural network (CNN) which avoids the usage of hand crafted features. CNN is a recent development which is a type of artificial neural network usually designed to extract features from raw pixels of images and then classifies the given images into a set of classes. CNN is designed specifically to reorganize two dimensional shapes with a high degree of invariance to translation, scaling, skewing and other forms of distortion. The structure includes feature extraction, feature mapping, pooling, and sub-sampling layers. CNN consists of a number of convolution and sub-sampling layers optionally followed by fully connected layers. Similar to the MLFFNN, back propagation algorithm is used to train the CNN model. Gradient descent algorithm can be used to optimize the parameters.

Convolution Neural Networks are biologically inspired variants of Multilayer Neural Network. From experiments mentioned in [5] it is known that the visual cortex of animals is a complex network of cells. Each cell is sensitive to small sub-region of the visual fields, known as the receptive field. According to [4], there are two kinds of cells: Simple cells and Complex cells, where simple cells extract features and complex cells combine several such local features from a spatial neighbour-hood. CNN tries to imitate this structure, by extracting the features in a similar way from the input space and then performing the classification, unlike the standard techniques where features are extracted manually and provided to the model for classification [3].

Based on the literature, Deep Neural Networks (DNN) for hand written digit classification gives better performance than existing methods. We have used a variant of DNN called Deep convolution Neural Networks (DCNN) for feature extraction and classification. Neural networks can be used for classification task as well as for feature extraction. So a separate feature extraction algorithm need not be used. Various configurations of DCNN are used for our experimental studies. Among different architectures that we have considered, the architecture with 2 levels of convolutional and pooling layers, followed by fully connected layers is used for the proposed hand written digit recognition task.

Handwritten digit recognition can be considered as a subtask of hand written character recognition. Hand written character recognition is a broad area where a given character is recognized automatically by a machine. The major challenges of hand written character recognition are: writing style and size of characters varies from person to person. Even for the same person the hand writing varies from time to time and it is highly influenced by the context. That means the style of writing a character is certainly depends on the previous characters and the characters that follows. The style of the characters also depends on the speed of writing.

With the advances in machine learning algorithms the success of handwritten character recognition is improved. In this task we have considered hand written digit recognition, a subtask of hand written character recognition task as the applications of this task are plenty. Amount identification in case of Bank cheques, recognizing zip codes in case of postal letters are the two most important real time applications of hand written digit recognition. Recent literature shows that performance of Convolutional Neural Network (CNN) is promising on images. We have used neural network based models for hand written digit recognition classification. The model is tested using the bench mark MNIST dataset.

II. EXISTING WORK

Character recognition is a fundamental, but most challenging in the field of pattern recognition with large number of useful applications. More precisely Character recognition is the process of detecting and recognizing characters from the input image. If the input image contains characters written by hand then it is called as hand written digit recognition. There are two types of Handwriting recognition: offline handwriting recognition and online handwriting recognition. If handwriting is scanned and then understood by the computer, it is called offline handwriting recognition. In case, the handwriting is recognized while writing through touch pad using stylus pen, it is called online handwriting recognition. The major challenges of this task are: No two persons in the world write exactly the same way. Same person writes in different styles. It is not guaranteed that the same person writes in the same way at different times.

In [3], an efficient handwritten digit recognition algorithm has been proposed based on HoG features and SVM is used for classification. A K-nearest Neighbor Classifier (KNN) based hand written digit classifier is proposed in [6]. The same is extended using [7] for Arabic handwritten digit recognition. A Persian/Arabic handwritten digit recognition is proposed in [8] that uses local binary patterns.

An approach that uses Alternately Trained Relaxation Convolutional Neural Network (CNN) is proposed in [9]. An algorithm that uses SVM based classification is proposed in [10] by Lauer etal. A dynamic time warping (DTW) based recognizer for handwriting and gesture recognition is proposed in [14]. A deep learning based algorithm is proposed in [15] and these models dramatically improved the then state-of-the-art performance. A Recurrent convolutional based Neural Network is proposed in [16]. In [17] a Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the 'extreme learning machine' algorithm. A Handwritten digit recognition that uses support vector machine optimized by bat algorithm is proposed in [2]. A cost effective isolated handwritten recognition method is proposed in [1] Bangla character and digit recognition.

III. PROPOSED MODEL

The objective of our task is to assign a label to the given handwritten digit. The impressive gain in performance obtained using deep neural networks (DNN) for various tasks encouraged us to apply DNN for hand written digit classification task. We have used a variant of DNN called deep convolutional Neural Networks (DCNN) for feature extraction and image classification. Neural networks can be used for classification as well as for feature extraction. Our whole work can be better seen as two different tasks that are elegantly combined into a single task. In the first task, features are extracted automatically and then they are fed to model to classify the given features into one of the appropriate classes. These classes correspond to the digits in our database. We have considered different architectures for our experimental analysis. Among the different architectures that we have considered, the architecture with 2 levels of convolutional and pooling layers, followed by a fully connected output layer is the best one. In this task the automatically learned features extracted by DCNN are fed to fully-connected feed forward neural network for classification. Experimental studies show that the performance of using automatically learned features are better than handcrafted features. Based on our experiments we claim that performance of hand written digit classification is significantly better than the results of the other models that use handcrafted features followed by classification.

A convolutional neural network (CNN) is a type of artificial neural network usually designed to extract features from data and to classify given high dimensional data. CNN is designed specifically to reorganize two dimensional shapes with a high degree of invariance to translation, scaling, skewing and other forms of distortion. The structure includes feature extraction, feature mapping and sub-sampling layers. A CNN consists of a number of convolutional and subsampling layers optionally followed by fully connected output layers. Back propagation algorithm is used to train the model.

For illustration consider a 28x28 grey scale images where each image is represented as a gray scale value and can be represented as a 784 dimensional vector, using the raw pixels to represent image. Considering color images, different values of color (R, G and B), the same image can be better represented as a vector of size 2252 (dimensions). To model such high dimensional data using shallow networks involves estimation of large number of parameters. Unless training data is large for such models the lead to over fit the data.

Convolutional Neural Network can handle such problems by leveraging the ideas of local connectivity, parameter Sharing and Pooling/ Subsampling and fully connected layers. Each convolutional module consists of 2 sub modules called convulutional layer followed by pooling layer as shown in the Figure1. Convulution layer, can extract different representations of the image as a feature map where each feature map represents one specific set of type of feature of the image. Pooling layer on the other hand is used to reduce the redundancy generated in the convolution layer.

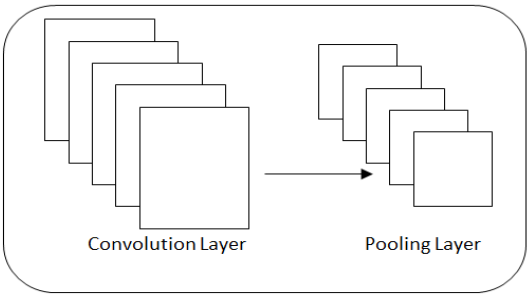


Figure 3: Convolution module.

a. Convolution layer: The input to this layer is raw pixels of an image and output is different feature representations of the same image called feature maps. The input image is divided into number of blocks which are

usually equal in size. A filter of size k*k is applied on each block of the image and convolved with same filter applied on all the blocks of the image. The output generated by this process is called as a feature map as it represents a specific feature of the image. Different feature maps can be obtained by convolving all the blocks of the image with different filters. The blocks can be either overlapping or non-overlapping in nature. If the blocks are overlapping then they share some common part of the image and non-overlapping blocks do not share any. In order to extract smooth features, overlapping blocks are being considered. In this way local features are extracted and the exact location of the features is of no importance. This is beneficial until the relative location with respect to other features is preserved. Figur2 illustrates an example of local connectivity.

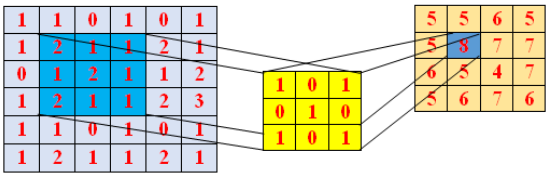


Figure 4: Local connectivity

In Figure2 the input is of size 6x6 whereas the output is of size 4x4. To preserve the same size in the output padding can be used. Another disadvantage of convolution without padding is corner pixels involve in very few convolutions whereas the middle pixels involve in more convolutions. This effect can also be avoided by using padding.

Parameter Sharing: Each computational layer, also known as convolution layer in the network is allowed to have multiple filters and each filter results in a different feature map. The neurons within a feature map are constrained to share the same weights. This constraint is to reduce the number of parameters in the network. If different filter is used for different blocks within the same feature then the number of parameters grows by a larger extent. To avoid this same filter is applied on all the blocks of the image within the feature map. This guarantees reduction in parameters and shift invariance.

Figure3 helps us to clearly understand the process of parameter sharing: Each hidden unit in a feature map is connected to different blocks of an image and extracts same type of features. Hidden units in different feature maps extract different features from the same block. For example one filter is useful to extract edges and other filter may be useful to extract color related features and some other filter may be used to extract lighting conditions.

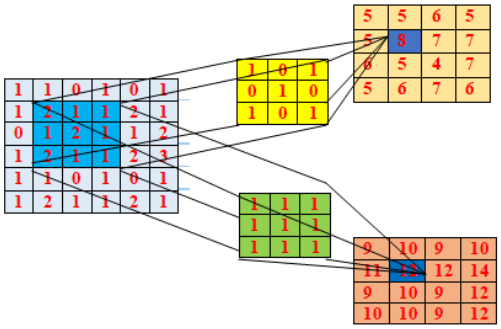


Figure 5: Parameter sharing

Pooling and sub-sampling: The objective of this step is to remove the redundant information in the image in other words to reduce the size of the input. Different types of pooling can be applied namely max pooling, min pooling and average pooling. A window of some predefined size is selected in both the methods. In Max pooling, maximum activation value among the values in a window is considered. In Average pooling average of the activation values in a window is considered and in min pooling, minimum activation value among the values in a window is considered.

Proposed Method:

Figure5 illustrates the detailed architecture of CNN for classification of 28x28 handwritten character images. This architecture comprises of 2 convolutional layers followed by a pooling layer

	Layers	Configuration	Accuracy
1	C-P-C-P-FC	Conv1: 3X3, stride:1, padding: same, Activation: ReLU Feature Maps: 32 Conv2: 3X3, stride:1, padding: same, Activation: ReLU Feature Maps: 64 Initialization: Xavier, Optimizer:Adams, Loss: crossmax with softmax learning_rate = 0.001, epochs = 15, batch_size = 100	98.84
2	C-P-C-PC-P-FC	Conv1: 3X3, stride:1, padding: same, Activation: ReLU Feature Maps: 32 Conv2: 3X3, stride:1, padding: same, Activation: ReLU Feature Maps: 64 Conv2: 3X3, stride:1, padding: same, Activation: ReLU Feature Maps: 128 FC: 625-10 Initialization: Xavier, Optimizer:Adams, Loss: crossmax with softmax learning_rate = 0.001, epochs = 15, batch_size = 100	99.38

which are finally connected to a fully connected to output layer. In the first convolutional layer receptive field is a 3x3 block, and each image is composed of such multiple blocks in an overlapping fashion. Thus each image is described in terms of 26x26 overlapping receptive fields. A convolution layer consists of several feature maps, here the first convolution layer consists of 4 feature maps. The neurons in the same feature map are constrained to have the same weights to reduce the number of parameters to be estimated.

The next layer is the sub-sampling layer with a window of size 2x2. Input to this layer is a 4x26x26 matrix and it is reduced to 4x13x13. Blocks at this stage are non-overlapping. The reduction in the size of the matrix can be performed either by average pooling or by max pooling [3] [4]. The subsequent layers are two convolutional and sub-sampling layers with an increase in the number of feature maps and decrease in the spatial resolution. The output layer is a fully connected layer. Output of this layer gives features of the input image. Further classification of these features into classes can be done using fully connected neural network.

IV. EXPERIMENTAL RESULTS

In this section the model that gives best performance is considered and results are reported.

Summary of the Dataset: The task is to classify hand written characters into one of the given 10 classes labelled from 0 to 9. We have considered MNIST digit dataset for our experimental results. The dataset consists of 70,000 different handwritten characters in different styles. The dataset is divided into training and test. Training partition consists of 60,000 and test set contains 10,000 images. All the characters are gray colored images. Each image is of size 28x28.

Classes	Train	Test	Total	Dimension
10	60,000	10,000	70,000	28x28

Table1: Data set after Summary

Model Selection: Raw image pixels are used images are given as input to the CNN model for training. Cross validation method is used to arrive at the best model. In the experiments we have tried different architectures and the model that gives the highest performance is considered as the best model for the given dataset.

Proposed CNN architectures: The Convolutional Neural Network is trained using Adam's optimization and Stochastic Gradient Descent with Momentum. The network consists of an input layer, followed by three convolutional and average pooling layers and followed by a soft max fully connected output layer to extract features. After extracting features, 2 layer hidden neural-network is used for classification. We used different configuration setups to extract features from the given data. In all these configurations pooling layer window size is fixed as 2x2, and we tried with different epochs with a batch size of 100. Some of these setups are motivated by literature. Some of the guidelines we observed from the literature are: (i) Fewer convolution filters in the early layers and increased filters in the later layers. (ii)Larger window size early in the layers compared to the later layers (iii) Drop outs can help to avoid over-fitting.

Choice of convolution window size, number of filters at different layers, activation function used at each layer and pooling type (average or max pooling) are selected based on the cross validation method.

Table2 shows different configurations used in the experimental studies:

Comparison of deep features with hand crafted features: In this section, the effect of deep features is compared with that of hand crafted features.

Features	Classification model	Accuracy
Raw pixels	Linear Model	90.23%
	Neural Network	94.55%
Histogram	C-SVM	22.40%
	Nu-SVM	23.40%
HoG	C-SVM	96.70%
	Nu-SVM	96.00%
GIST	C-SVM	97.20%
	Nu-SVM	97.50%
Deep Features	Deep NN	97.25%
	CNN	99.38%

Table2: Comparison of accuracy using different features

Graph in figure4 shows how the error is reduced over every epoch when the model is trained using deep convolution neural network. Initially the error is very high but as the model is trained the error is reduced and results an accuracy of 99.38%.

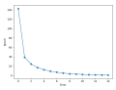


Figure6: Loss function of the model trained on MNIST

REFERENCES

1. Md Zahangir Alom, Paheding Sidike, Mahmudul Hasan, Tarek. Taha, and Vijayan K. Asari, " Handwritten Bangla Character Recognition Using the State-of-the-Art Deep Convolutional Neural Networks", 2018.
2. Tuba, Eva, Milan Tuba, and Dana Simian. "Handwritten digit recognition by support vector machine optimized by bat algorithm.", WSCG (2016).

3. Djork-Arne Clevert, Thomas Unterthiner & Sepp Hochreiter , “fast and accurate deep network learning by exponential linear units (elus)”, ICLR 2016
4. Jyostna Devi Bodapati et al, “A novel face recognition system based on combining eigenfaces with fisher faces using wavelets", Procedia Computer Science2 (2010): 44-51.
5. Jyostna devi Bodapati et al, “scene classification using support vector machines with lda”, Journal of Theoretical & Applied Information Technology. 2014 May 31;63(3).
6. Zhang, Hao, et al. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2. IEEE, 2006.
7. Al-Shalabi, Riyad, and Rasha Obeidat. "Improving KNN Arabic text classification with n-grams based document indexing." Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt. 2008.
8. Ilmi, Nurul, W. Tjokorda Agung Budi, and R. Kurniawan Nur. "Handwriting digit recognition using local binary pattern variance and K-Nearest Neighbor classification." Information and Communication Technology (ICoICT), 2016 4th International Conference on. IEEE, 2016.
9. Wu, Chunpeng, et al. "Handwritten character recognition by alternately trained relaxation convolutional neural network." Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014.
10. Mustafa S. Kadhm, Alia Karim Abdul Hassan, “Handwriting Word Recognition Based on SVM Classifier“, International Journal of Advanced Computer Science and Applications, Vol. 6, No. 11, 2015
11. Jyostna devi Bodapati et al, “An intelligent authentication system using wavelet fusion of K-PCA, R-LDA”, IEEE International Conference on Communication Control and Computing Technologies (ICCCCT), 437-441, 2010,
12. Pradhan, Debasish "Enhancing LBP Features for Object Recognition using Spatial Pyramid Kernel." (2017)
13. Harris, Samuel, et al. "LBP features for hand-held ground penetrating radar." Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXII. Vol. 10182. International Society for Optics and Photonics, 2017.
14. Hsu, Yu-Liang, et al. "An inertial pen with dynamic time warping recognizer for handwriting and gesture recognition." IEEE Sensors Journal 15.1 (2015): 154-163.
15. Cireşan, Dan Claudiu, et al. "Deep, big, simple neural nets for handwritten digit recognition." Neural computation 22.12 (2010): 3207-3220.
16. Visin, Francesco, et al. "Renet: A recurrent neural network based alternative to convolutional networks." arXiv preprint arXiv:1505.00393 (2015).