Classification Of E-MNIST Dataset Using Neural Networks

Aron Pereira
Machine Learning Intern
AI Technology and Systems
aronpereira20@gmail.com
www.ai-techsystems.com

***Abstract*—This paper presents an implementation using Neural Network to achieve the classification of the E-MNIST(Extended MNIST) handwritten digit database. Here, we use plot of train dataset size and classification accuracy to determine the performance of the neural network. The experimental results show that to a certain extent, the Convolutional Neural Network can be used to solve classification problem in the real world. Besides, the data in the dataset is processed to its best possible state. It does help to reduce the network size and thus increase the speed of the neural network, improving the performance. However, the accuracy rate cannot be guaranteed and potential reasons are discussed. At last, the results generated by our system are compared with those in another research paper tested the same E-MNIST handwritten digit database. It seems that the results of this experiment using CNN is better than the traditional machine learning algorithms.**

Keywords – Deep-Learning, Neural Networks, E-MNIST handwritten digit database

I. INTRODUCTION

Handwritten digit recognition is an important problem in optical character recognition, and it can be used as a test case for theories of pattern recognition and machine learning algorithms. To promote research of machine learning and pattern recognition, several standard databases have emerged. The handwritten digits are preprocessed, including segmentation and normalization, so that researchers can compare recognition results of their techniques on a common basis as well as reduce the workload [1].

The MNIST dataset remains the most widely known and used dataset in the computer vision and neural networks community. However, a good dataset needs to represent a sufficiently challenging problem to make it both useful and to ensure its longevity [3]. The MNIST database is a subset of a much larger dataset known as the NIST Special Database 19 [4]. This dataset contains both handwritten numerals and letters and represents a much larger and more extensive classification task, along with the possibility of adding more complex tasks such as writer identification, transcription tasks and case detection

This paper introduces such a suite of datasets, known as Extended Modified NIST (EMNIST). Derived from the NIST Special Database 19, these datasets are intended to represent a more challenging classification task for neural networks and learning systems. By directly matching the image specifications, dataset organization and file formats found in the original MNIST dataset, these datasets are designed as drop in replacements for existing networks and systems

The EMNIST dataset can be accessed and downloaded from https://www.kaggle.com/crawford/emnist

II. METHODOLOGY

As mentioned above, in this paper we use MNIST handwritten digit database in which the handwritten digits have been preprocessed including segmentation and normalization. This paper introduces CNN classifier to the classification tasks based on these datasets. The purpose of the classifiers is to provide a means of validating and characterizing the datasets whilst also providing benchmark classification results. The nature and organization of the EMNIST datasets are explained through the use of these classification results

A. Conversion Process

The conversion process transforms the 128 × 128 pixel binary images found in the EMNIST dataset to 28 × 28 pixel images with an 8-bit gray-scale resolution that match the characteristics of the digits in the MNIST dataset. The EMNIST dataset contains 814, 255 images in four different hierarchies which affect the labeling and organization of the data.

For this work, only the By Class hierarchy is used. In order to convert the dataset, each digit is loaded individually and the images are rotated in order to convert the inclined images. As the size and shape of the characters and digits vary both from class-to-class and from writer-to-writer, there is significant variance in the size of the region of interest. Whereas the original MNIST conversion technique down-sampled the digits to either a 20×20 pixel or a 32×32 pixel frame before placing it into the final 28×28 pixel frame, the technique used in this paper attempts to make use of the maximum amount of space available.

B. Training and Testing Splits

The handwriting data used in the Special Database 19 was collected from both Census employees and high-school students. The specifications provided alongside the dataset include a suggestion that the handwritten digits from the student corpus be used as the testing set. Although the argument that the student handwriting represents a harder task and therefore should be used as an unseen testing set has merit, it does raise questions as to whether there is enough similarity and consistency between the two sets of participants. For this reason, the original MNIST dataset uses a different training and testing split from the one specified and recommended in the user guide supplied alongside both dataset releases. The creation of the EMNIST dataset therefore follows the same methodology used in the original MNIST paper [1] in which the original training and testing sets were combined and a new random training and testing set were drawn. The resulting training and testing datasets thereby contain samples from both the high-school students and the census employees respectively are considered using trial and error method.

C. Validation Partitions

Many iterative training algorithms make use of a validation partition to assess the current performance of a network during training. This needs to be separate from the unseen testing set to maintain the integrity of the results. Instead of including a separate validation set for each class, the balanced datasets in the EMNIST dataset contain a specially balanced subset of the training set intended specifically to be used for validation tasks.

STRUCTURE AND ORGANIZATION OF THE EMNIST DATASETS.

Name	Classes	No. Training	No. Testing	Validation	Total
By_Class	62	697,932	116,323	No	814,255
By_Merge	47	697,932	116,323	No	814,255
Balanced	47	112,800	18,800	Yes	131,600
Digits	10	240,000	40,000	Yes	280,000
Letters	37	88,800	14,800	Yes	103,600
MNIST	10	60,000	10,000	Yes	70,000

The table above contains a summary of the EMNIST datasets and indicates which classes contain a validation subset in the training set. In these datasets, the last portion of the training set, equal in size to the testing set, is set aside as a validation set. Additionally, this subset is also balanced such that it contains an equal number of samples for each task. If the validation set is not to be used, then the training set can be used as one contiguous set.

D. Classifiers

Convolutional Neural Networks are a special kind of multi-layer neural networks. They are also trained with the backpropagation algorithm but with the different architecture from other neural networks. As Y. LeCun introduced, Convolutional Neural Networks are designed to recognize visual patterns directly from pixel images with minimal preprocessing and they can recognize patterns with extreme variability such as handwritten characters even digits (Yann.lecun.com, n.d.).

.

III. RESULTS

The EMNIST By Merge an EMNIST By Class datasets both contain the full 814,255 characters contained in the original NIST Special Database 19. The primary differences between the two datasets are the number of classes, the number of samples per class, and the order in which the characters appear in the dataset. The By Class dataset contains the full 62 classes comprising 10 digit classes, 26 lowercase letter classes, and 26 uppercase letter classes.. The number of characters per class varies dramatically through dataset, as is clearly visible in Figure 1. The dataset also contain vastly more digit samples than letter samples.

A CNN classifier was used to explore the nature of the classification tasks designed around the dataset. The full training set was used to train the classification networks with one hidden layer, which was subsequently tested on the testing set. Ten trial of the experiment were performed and the results of the classifiers are shown in Figure 2. The graph shows the accuracy vs dataset size for various training sizes of dataset and the test size was kept constant. The results show that as the dataset size increased, the accuracy of the classifier increased

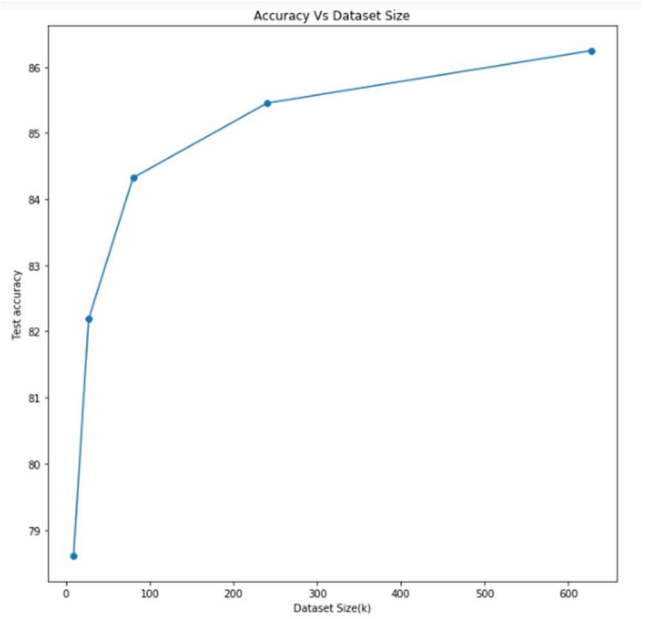

Fig 2: Dataset size vs Accuracy graph

The classifier achieved a maximum accuracy of 86.24% when the dataset size was 697k and a minimum accuracy of 78.6% when the dataset size was 9k. The EMNIST dataset also contains a slightly different set of digits from the original MNIST dataset, primarily due to the need to balance the number of samples in each class. Although this may have impacted the accuracy achieved on the dataset somewhat, it does not explain the significant and consistent improvement in accuracy achieved using the EMNIST digits.

IV. CONCLUSION

This paper introduced the EMNIST datasets, a suite of six datasets intended to provide a more challenging alternative to the MNIST dataset. A comparison of the performance of the classification task on a subset of digits against the original MNIST dataset served to further validate the conversion process. The EMNIST datasets therefore provide a new classification benchmark that contains more image samples, more output classes, a more varied classification task, and a more challenging classification task than MNIST whilst maintaining its structure and nature. The classifier built was way better than the traditional MNIST dataset and achieved a better accuracy. This therefore represents a new and modern performance benchmark for the current generation of classification and learning systems.

ACKNOWLEDGMENT

I would like to thank the mentors of AI Technology and Systems,at which the most steps of this work was completed.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2323, 1998.

[2] A. Krizhevsky, “Learning Multiple Layers of Features from Tiny Images,” Ph.D. dissertation, University of Toronto, 2009.

[3] G. Orchard, A. Jayawant, G. K. Cohen, and N. Thakor, “Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades,” Frontiers in Neuroscience, vol. 9, pp. 1–15, nov 2015. .

[4] P. Grother, “NIST special database 19 handprinted forms and characters database,” National Institute of Standards and Technology, Tech. Rep., 1995.

[5] L. Wan, M. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, “Regularization of neural networks using dropconnect,” Icml, no. 1, pp. 109–111, 2013.

[6] D. Cirean, U. Meier, and J. Schmidhuber, “Multi-column Deep Neural Networks for Image Classification,” International Conference of Pattern Recognition, no. February, pp. 3642–3649, 2012.

[7] I. Sato, H. Nishimura, and K. Yokoi, “APAC: Augmented PAttern Classification with Neural Networks,” Arxiv, 2015.

[8] J.-R. Chang and Y.-S. Chen, “Batch-normalized Maxout Network in Network,” Arxiv, 2015