Failure of deep-learning models in generalising pattern recognition tasks

Krishna Chaitanya Reddy Tamataam
Department of Mathematics
IIT Delhi
Delhi 110016, India
mt6180785@maths.iitd.ac.in

Abstract

Neural networks have gained popularity in the present day for their power to deduce complex patterns from the data given. But the problem with neural networks is that they are too complex and we have a very limited understanding of how the model predicts. Usually we want the predictions to be more generalised across different datasets, but neural network often tend to show high dataset dependency. In this paper we demonstrate how the accuracy change on slightly altering the dataset and how the model performs on new dataset. Also we would be touching some theories regarding why humans might excel in such tasks.

1 Introduction

In the recent times, due to availability of large amount of data and high computational power, Artificial Intelligence (AI) and Machine Learning (ML) algorithms play a very important role in automation. From basic neural networks to predict a simple non-linear functions to speech recognition using Recurrent Neural Network (RNN), there are several applications of AI that we use in our day-to-day lives. In today's world these models have been trained rigorously and have achieved high accuracy. But the Big Mystery with Deep Learning models is that no one really understand what is happening inside the model. Often it is referred to as a black box and it is difficult to explain the results predicted by these models. Also deep-learning models often tend to be dataset dependent. This tells that model is not learning what we want it to learn and recognise, rather learning the idiosyncrasies in that particular dataset. Humans on the other hand perform the task of pattern recognition robustly across different datasets as we try recognise and learn the important features quickly for prediction and we actually have a definitive way to explain our prediction. In this paper we take up the task of handwritten digit and train a basic CNN network on the MNIST dataset. Then we slightly modify the test dataset and observe the change in accuracies, also we check the accuracies on the USPS dataset that is converted into MNIST format and observe the change in accuracy. This suggests that neural networks have a long way to catch up to reach the human level in identifying patterns and we cannot yet completely rely on the expertise of neural networks, especially in the field of medical diagnosis as it cannot explain its predictions correctly.

2 Datasets

MNIST Dataset is a very famous dataset that has grayscale images(28*28) of handwritten digits. The dataset contains of 60,000 labelled images for training and 10,000 images for testing with similar proposition of images for each label.This is necessary to ensure a similar distribution between training and test dataset. We also used the USPS(United States Postal service dataset) which consists of total of 9,298 16*16 pixel grayscale samples, the images are centered, normalized and show a broad range of font styles.

3 Model

3.1 Architecture

We used a very simple convolution neural network(CNN) model which has a 2-d convolution layer with 32 filters of size 5*5. On this output we applied a max-pooling layer of size 2*2, Then added a dropout layer to this layer with a probability of 0.2 for regularisation. Then we added two more dense layers with activation's of relu and softmax respectively.The softmax layers returns a vector that represents the probability of each label given the image.

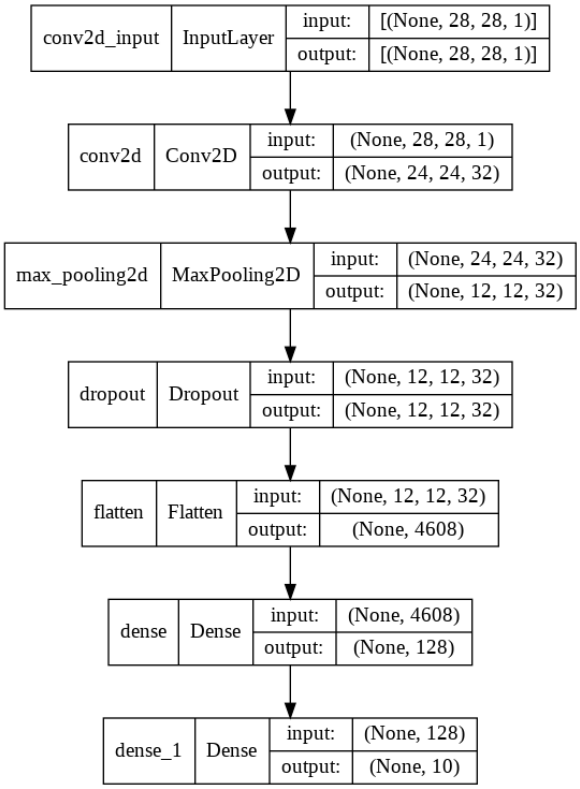

Figure 1: Architecture

3.1.1 Training and optimisation

We used the simple *model.fit()* for the training loss which minimizes the categorical-cross entropy loss function through Adam optimizer. We have used the mini-batch gradient descent algorithm with a batch-size of 200 and trained the model for 10 epochs. The training error was around 99.53 while the test accuracy was about 99.06.we have plotted the train and test loss over epochs for training as well as the confusion matrix

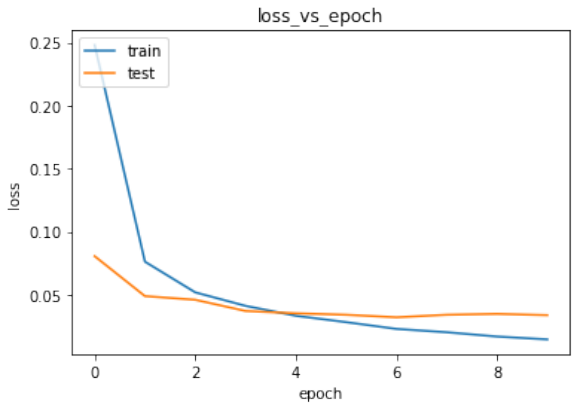

Figure 2: loss over epochs

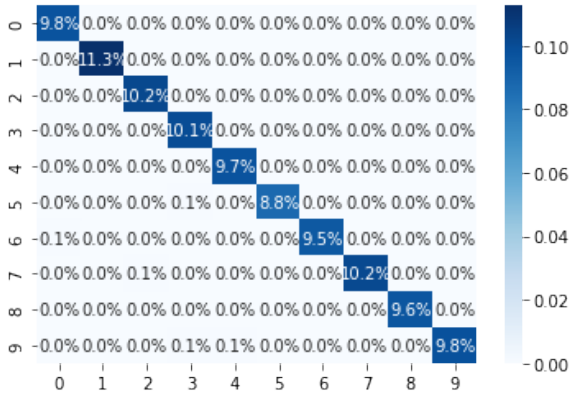

Figure 3: Confusion Matrix

4 Experiment

4.1 Preprocessing

First the input image is a 8-bit gray scale image of size 28*28. first we convert the input from uint8 to float32 format and then divide the input by 255 so that all input values lie between [0,1]. This is done so that all input values lie in the same range and every pixel is assumed to have equal importance during the learning and we can expect a uniform learning curve.

4.2 Settings

We perform the following alterations to the test dataset and observe the difference in accuracy.

- We introduce salt and pepper noise in the image with different probability till 0.2. After that it becomes difficult even for humans to recognise it.
- We rotate the image till 45 degrees by keeping a difference of 5 degrees.
- We add rotation followed by adding noise to the image.
- We convert the USPS dataset into MNIST format and then compare the accuracy dip.Since USPS dataset is a 16*16 image, we resize the image into 28*28 through linear interpolation.

We are using average accuracy as a metric to observe the difference between the altered datasets.Other metrics that are available for each label are precision, recall, F1-score and confusion matrix are provided in the experimental logs at the end of the report.

4.3 Results

We use the following notations to define the performance metrics we use in this report:

- **TP** (True Positive): Model predicts Positive and it is actually Positive.
- **FP** (False Positive): Model predicts Positive but it is actually Negative.
- **FN** (False Negative): Model predicts Negative but it is actually Positive.
- **TN** (True Negative): Model predicts Negative and it is actually Negative.

Now, we define the performance metrics of our interest.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

noise	accuracy	precision	recall	F-score
0	0.9901	0.9901	0.9901	0.9901
0.05	0.9161	0.9261	0.9161	0.9147
0.1	0.726	0.815	0.726	0.7048
0.15	0.5362	0.7319	0.5362	0.487
0.2	0.379	0.6814	0.379	0.3311

Table 1: noise vs accuracy for the test dataset

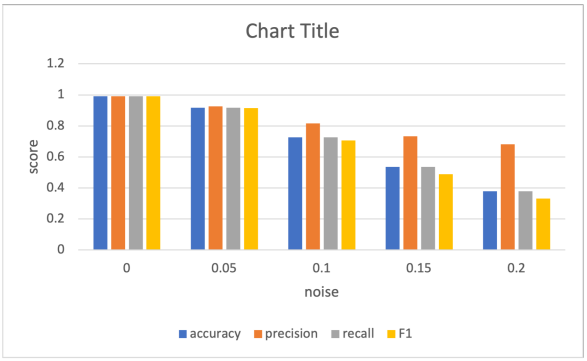

Figure 4: scores over noise

For the USPS dataset which contains approximately 9,298 images the accuracy is 0.7898, precision is 0.8184, recall is 0.7898, F1-score is 0.7666.

angle	accuracy	precision	recall	F-score
5	0.9877	0.9877	0.9877	0.9877
10	0.9818	0.9819	0.9818	0.9818
15	0.9708	0.971	0.9708	0.9707
20	0.9503	0.951	0.9503	0.9503
25	0.9048	0.9064	0.9048	0.9046
30	0.837	0.8402	0.837	0.8367
35	0.746	0.7529	0.746	0.7452

Table 2: rotation angle vs accuracy for the test dataset

This clearly is a significant difference and tells that the model is holding on to specific features on the dataset rather than learning a generalised pattern. In the present day people are just focusing on accuracy of the model rather than its interpretability(i.e how it is making a decision). This makes deep learning models unreliable in medical domain.

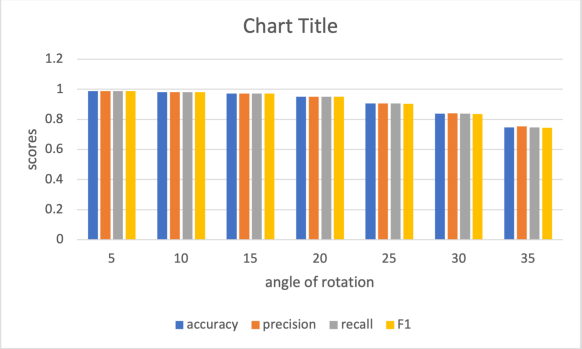

Figure 5: Scores over angle of rotation

angle	accuracy	precision	recall	F-score
5	0.8972	0.9104	0.8972	0.8953
10	0.8911	0.9032	0.8911	0.8896
15	0.8676	0.8826	0.8676	0.8664
20	0.82	0.8412	0.82	0.8183
25	0.7631	0.7897	0.7631	0.7613

Table 3: rotation of rotation along with noise d = 0.05 vs accuracy for the test dataset

angle	accuracy	precision	recall	F-score
5	0.6937	0.7927	0.6937	0.6709
10	0.6842	0.7802	0.6842	0.6622
15	0.666	0.7638	0.666	0.643
20	0.6249	0.7199	0.6249	0.6002
25	0.5782	0.676	0.5782	0.5521

Table 4: rotation of rotation along with noise d = 0.05 vs accuracy for the test dataset

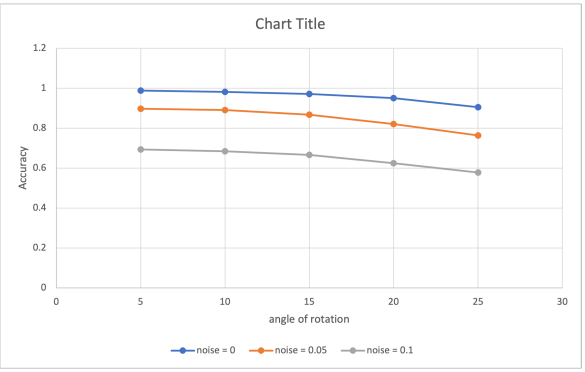

Figure 6: accuracy over angle of rotaion under different noises

4.4 Observations

- As noise in the image increases the accuracy decreases as expected.The accuracy appears to be decreasing exponentially with the noise.

- On adding noise to the images we can see(from figure 4) that the precision is getting maintained in the noise while accuracy, recall and F1 score are falling down.this says that the false positive results are considerably low compared to the false negative results on adding the noise.

- For small angle of rotation doesn't effect the prediction much but the dip in accuracy starts from an angle of 25 degrees(figure 5).

- On adding noise to the image this dip can be observed from even less 15 degrees angle(shown in figure 6).

5 Conclusion

5.1 Why Humans overestimate the contributions of AI

Eliza effect refers to the general inclination of people that we interpret simple outputs given by machines through algorithms to much higher meaning. Take a classical rule based algorithm that replies to people's typed statements. Many of the people their thought that the machine has emotions and empathy. Similar phenomena happens in the present day ML algorithms. They try to deduce simple rules from the data and we come under the effect that they are successful and smart.

5.2 Why humans perform better in recognition tasks

Russian Researcher Mikhil Bongard on his research on pattern recognition concludes that perception goes ways beyond simple pattern matching. He says that perception is far more than recognition of members of already existing categories, it involves spontaneous manufacture of new categories at arbitrary levels of abstraction.This ability helps humans actually to perceive things in different atmosphere. According to some other hypothesis Human mind is a collection a huge signal processing unit with a knowledge graph and a feedback loop. But no one know what processing is exactly happening in the brain and how the brain is storing it. It is very difficult even to design experiments for such hypothesis.

6 Fairness of the experiment

This experiment is not completely fair to the AI model when we are comparing it to the human level of accuracy. Humans have been training since childhood in the task of pattern recognition, starting from recognising faces to recognising complex patterns, to filtering out useless signals and allowing only certain signals to pass through. even on a different dataset humans could perform better due to this technique.AI in present day is just an algorithm without any mind of its own, but it might evolve in the future.

7 References

- CNN Model for Image Classification on MNIST and Fashion-MNIST Dataset,Kadam, Shivam and Adamuthe, Amol and Patil, Ashwini,2020.

- On the Brittleness of Handwritten Digit Recognition Models, Seewald Alexander,2009.

- Pattern Recognition, Mikhail Moiseevich Bongard, 1970.

- Liu, L., Nakashima, K., Sako, H., Fujisawa, H. Handwritten digit recognition: benchmarking of state-of-the-art techniques in pattern recognition.

- Digits - A Dataset for Handwritten Digit Recognition ,Alexander K. Seewald