Robotic Inference

Muhammad Khalaf

Abstract—The ”Robotic Inference” project is aimed at using the power of neural networks to classify objects from two different datasets. Nvidia Digits workflow is used to help accomplish two tasks. In the first part of the project, The well-known network architecture GoogLeNet, is used to classify images contain bottles, candy boxes and nothing, with an accuracy over 75% and an inference time of less than 10 ms. And in the second part, the same network is used to classify different paper currencies. The power of deep neural networks can now help blind people and those with vision impairment in many ways in their daily life.

1 Introduction

Artificial neural networks can be most adequately characterised as computational models with particular properties such as the ability to adapt or learn, to generalise, or to cluster or organise data [1].

Neural networks are used extensively in applications that require some sort of classification. Using the right learning algorithm, a multi-layered neural network can classify high dimensional data inputs into finite number of classes or clusters.

Neural Networks are applied to solve a wide variety of problems in science, engineering, finance, medicine etc.

They can be used to recognize individual writing styles, Google uses neural networks for image tagging, and Microsoft has developed neural networks that can help convert spoken English speech into spoken Chinese speech. Researchers at Lund University and Skne University Hospital in Sweden have used neural networks to improve long-term survival rates for heart transplant recipients by identifying optimal recipient and donor matches [2].

In this project , ANN is used to accomplish two tasks.

1.1 Task 1

The first task is to classify objects in a conveyor system, the neural network should classify an image as a bottle, candy box, or an empty image.

1.1.1 Task 2

And the second task is to classify different paper currencies, to help vision impaired people. Have that said, according to the World Health Organization, an estimated 253 million people live with vision impairment: 36 million are blind and 217 million have moderate to severe vision impairment [3]. The neural network is required to differentiate between a 200 Egyptian Pounds paper, a 5EGP paper and an invalid input.

2 Background / Formulation

The neural network architecture used in both tasks is the GoogLeNet architecture, which was the winner of the ILSVRC 2014 competition.

GoogLeNet introduced the concept of inception in CNNs. Inception is a novel technique that uses a composition of

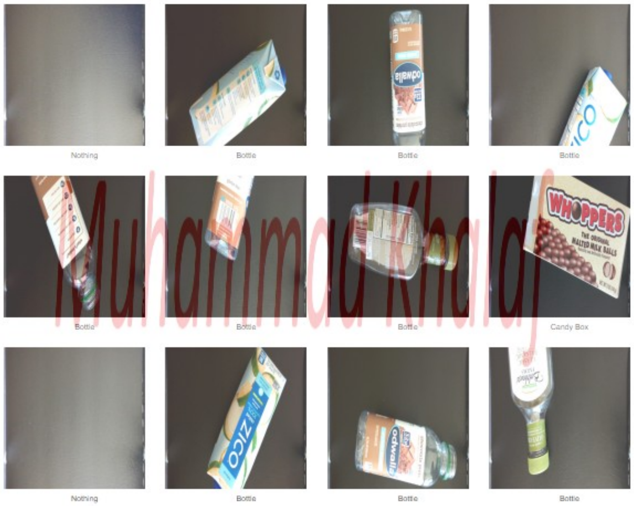


Fig. 1. Sample images from the first dataset

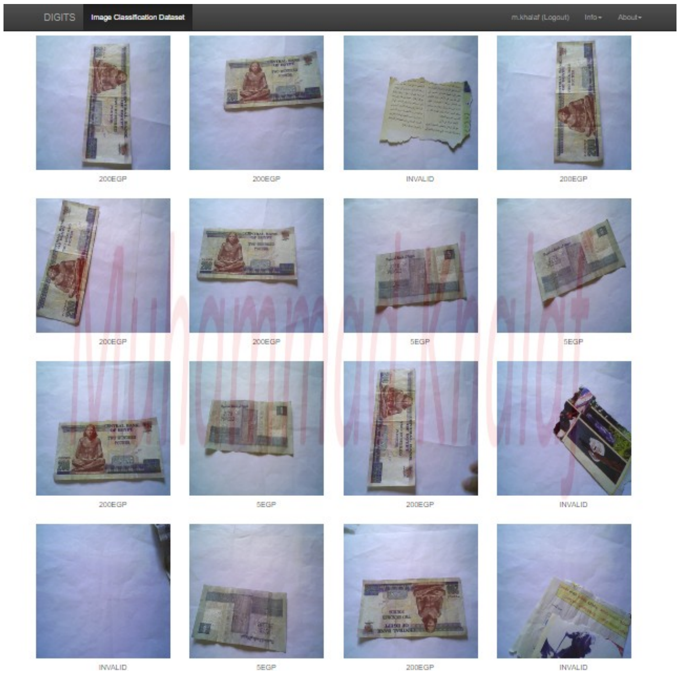


Fig. 2. Sample images from the second dataset

different convolution layers with a concatenated output instead of a simple convolution layer. This idea helped in reducing the number of parameters of the network, though it has many deep layers of convolution, max pooling, average pooling and inception module, and at the same time the results are much better [4].
Another advantage of GoogLeNet over other architectures is the reduced number of parameters. It uses 12x fewer parameters than AlexNet while being significantly more accurate [5]. Not to mention that a large number of parameters makes the network prone to overfitting. However, considering the inference time as a another metric, AlexNet architecture wins.
Different hyperparameters were used and the they were chosen using trials and errors. The learning rate starts with 0.01 and decays over time 3. A total number of 50 epochs are performed in the first task and 30 epochs for the second one, with a batch size of 64.

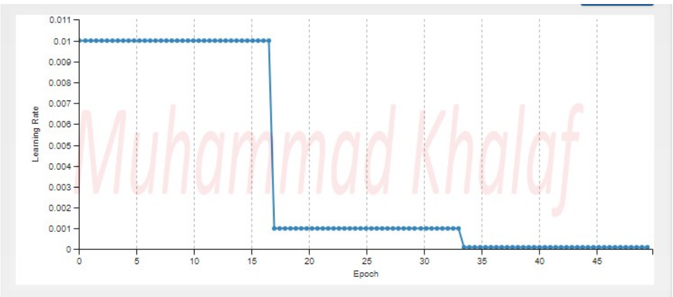


Fig. 3. learning rate curve

3 Data Acquisition

3.0.1 *Task 1*

The first dataset is provided by Udacity. It contains about 10 thousand images of bottles, candy boxes and a third class of images with no objects.

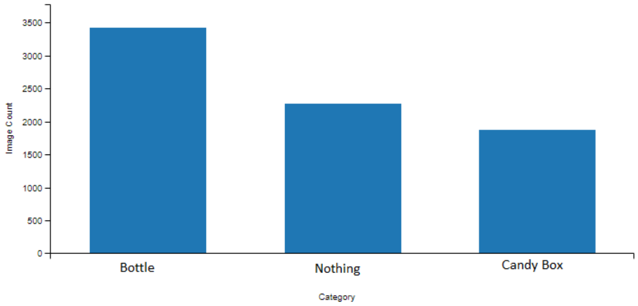


Fig. 4. Number of images of the first dataset used in training

3.0.2 *Task 2*

For the second task, the inference project idea, three different classes are introduced. Figure 2 shows samples of the data collected. There are many different characteristics between these classes. They are different in size, color and style. The Invalid images were a mix of magazine papers and empty images. Because of having a variety of characteristics, more images for the third class were collected. Figure 5 shows the number of images for each class used in training. With a total of 1050 images 75% of them were used in training and 25% were used in validation.

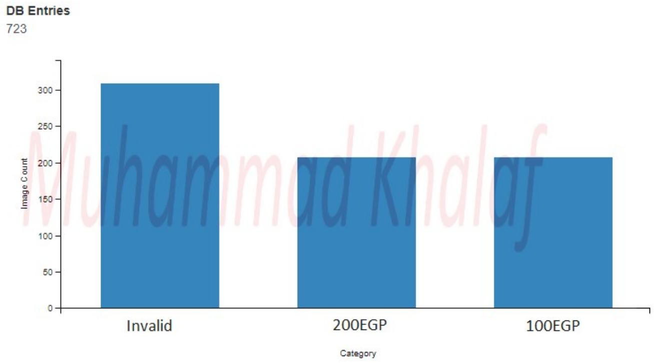


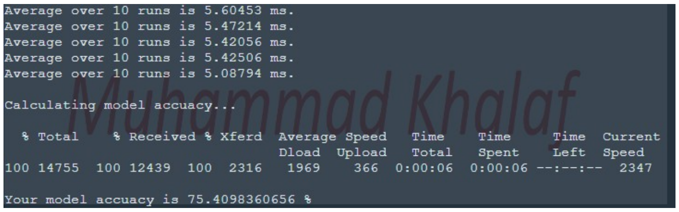
Fig. 5. Number of images of the second dataset used in training

These images were taken from an inexpensive webcam and using a simple opencv python script that uses the cv2.VideoCapture class to capture the images from the webcam.

4 Results

4.1 Task 1

The results for the first task using GoogLeNet architecture were good enough to pass the requirements. It achieved an accuracy of 75.4% and inference time of about 5.5 ms. Figure 6 shows the evaluation of the model, and figure 7 shows the accuracy and loss curves.



```
Average over 10 runs is 5.60453 ms.
Average over 10 runs is 5.47214 ms.
Average over 10 runs is 5.42056 ms.
Average over 10 runs is 5.42506 ms.
Average over 10 runs is 5.08794 ms.

Calculating model accuacy...

  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 14755  100 12439  100  2316   1969    366  0:00:06  0:00:06 --:--:--  2347

Your model accuacy is 75.4098360656 %
```

Fig. 6. Accuracy and inference time for the first task

4.2 Task 2

Again using the same network the network was able to classify most of the images correctly. Some results are shown in figures 8 and 9, and the accuracy and loss curves are shown in figure 10.

5 Discussion

As shown in the results section, the same network architecture was used in both tasks and was able to classify most of the images. To experiment different other architectures, AlexNet was also used but it gives a less accurate results.
However, in the second task, the currency paper

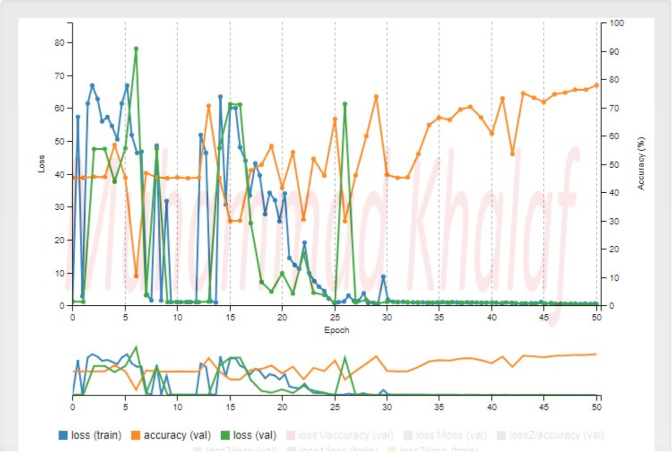
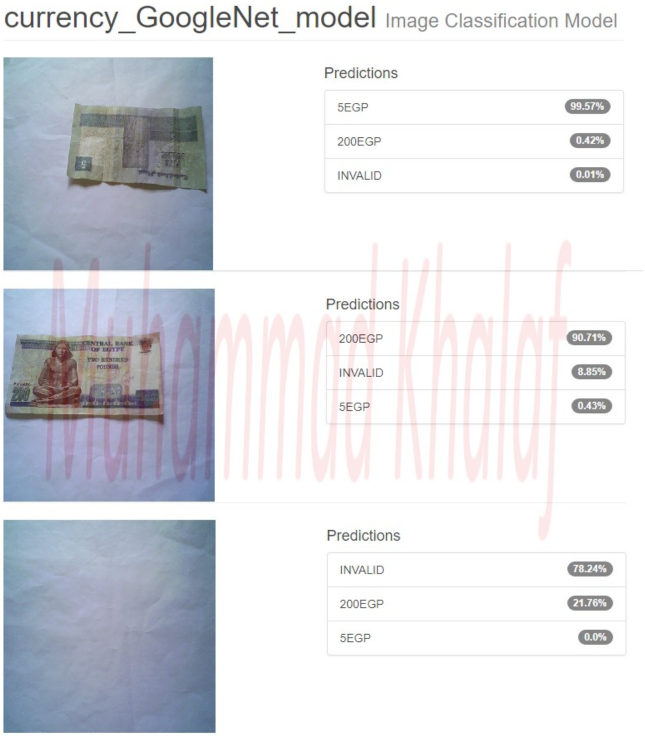


Fig. 7. Accuracy and loss curve for the first task


currency_GoogleNet_model Image Classification Model

Predictions

5EGP	99.57%
200EGP	0.42%
INVALID	0.01%

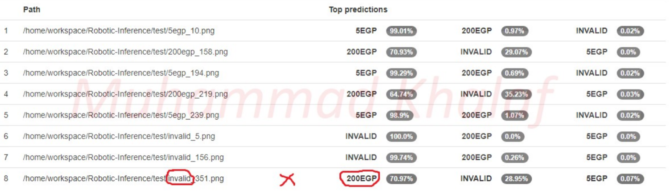
Predictions

200EGP	90.71%
INVALID	8.85%
5EGP	0.43%

Predictions

INVALID	78.24%
200EGP	21.76%
5EGP	0.0%

Fig. 8. task2 results



	Path	Top predictions					
1	/home/workspace/Robotic-Inference/test/5egp_10.png	5EGP	99.01%	200EGP	0.97%	INVALID	0.02%
2	/home/workspace/Robotic-Inference/test/200egp_158.png	200EGP	70.93%	INVALID	29.07%	5EGP	0.0%
3	/home/workspace/Robotic-Inference/test/5egp_194.png	5EGP	99.29%	200EGP	0.69%	INVALID	0.02%
4	/home/workspace/Robotic-Inference/test/200egp_219.png	200EGP	64.74%	INVALID	35.23%	5EGP	0.03%
5	/home/workspace/Robotic-Inference/test/5egp_239.png	5EGP	98.9%	200EGP	1.07%	INVALID	0.02%
6	/home/workspace/Robotic-Inference/test/invalid_5.png	INVALID	100.0%	200EGP	0.0%	5EGP	0.0%
7	/home/workspace/Robotic-Inference/test/invalid_156.png	INVALID	99.74%	200EGP	0.26%	5EGP	0.0%
8	/home/workspace/Robotic-Inference/test/invalid_351.png ✗	200EGP	70.97%	INVALID	28.95%	5EGP	0.07%

Fig. 9. A test shows a misclassification, the network is confused between the 200EGP images and Invalid images.

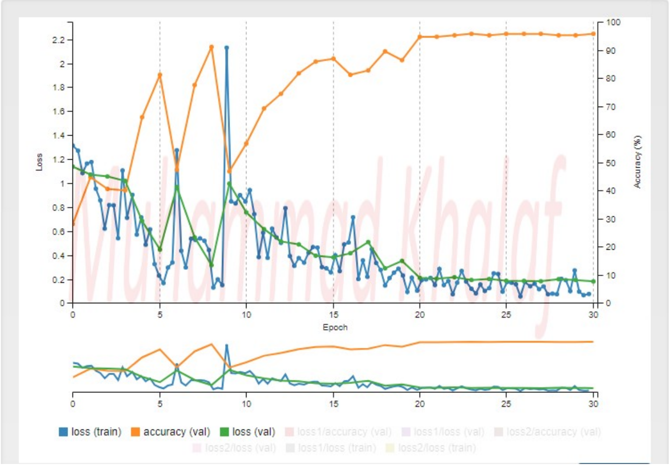


Fig. 10. Accuracy and loss curve for the second task

classification task, the results show that the network was not sure enough when it is given an Invalid image or an image of a 200EGP. This is certainly due to the lack of details and the small number of images given to the network to be trained on, about 300 images for the Invalid class and only 200 images for the 200EGP class.

Inference time does not actually matter for the currency recognition application, accuracy on the other hand is the main concern.

6 Conclusion / Future work

The idea of currency paper classification puts high constraints on the accuracy that should be achieved, and the results shown are not good enough. The dataset should be bigger, images should be of a higher quality and the network should be trained more and more to achieve better results.

This idea could be exploited in a mobile application that helps people with visual impairment. And of course it should contain more other currencies of different countries. Additioonally, Egyptian currency papers have two different faces. So for example, a 5EGP could be separated into two classes: 5EGP_FACE1 and 5EGP_FACE2.

A new personal network architecture may also be introduced, instead of using the GoogLeNet architecture, to help accomplish the task more accurately.

References

[1] K. Ben and v. d. S. Patrick, *An Introduction to Neural Networks*. 1996.

[2] Hagan, Demuth, Beale, and D. Jesus, *Neural Network Design, 2nd Edition*.

[3] W. H. Organization, “Blindness and visual impairment.” http://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[4] Deeplearning.ai, “C4w2l06 inception network motivation.” https://www.youtube.com/watch?v=C86ZXvgpejM.

[5] C. Szegedy et al, “Going deeper with convolutions.” https://arxiv.org/abs/1409.4842.