Peer review report
on

Gesture Recognition Based on Deep Learning

- Runlin Hou, Sifan Yuan, Yuxiang Song, Haocong Wang

1. Report Summary

The aim of this project is to implement recognition of hand gesture of numbers using deep learning frameworks ResNet and LeNet. Two datasets are used for the training and testing of the frameworks. They are:

a. Sign language MNIST dataset
    i. Training and Test cases consist of 0-25 representing A-Z
    ii. Each image is of size 28x28 , with grayscale levels of 0 -255
    iii. 27455 training datasets are used
    iv. 7172 test datasets are used
b. Leap motion dataset with corresponding depth map and color images from Kinect
    i. gestures representing numbers 1- 10
    ii. 1400 data samples are used

Two CNN models - LeNet5 and ResNet are used to implement gesture recognition.

LeNet5:

LeNet5 is a simple Convolutional Neural network. LeNet is trained only on gray scale images due to its simple architecture. The LeNet architecture is composed of Input image of size 32x32x1→Convolutional Layer: 6 filters of size 5x5 gives output images of size 28x28x6 →Pooling layer : reduces the size of each feature map. Can be average pooling or maximum pooling. This layer gives output of 14x14x6 images →Fully connected Layer→ Output Layer of 10 for recognizing the numbers.

LeNet implementation:

In this project, each of the 7 layers is implemented using PyTorch. Since the input images are RGB, modifications have been made to handle all the three color channels. Modifications are also made to accommodate two datasets from Kinect and Leap motion.

ResNet:

Deeper classical CNNs do not perform well as the depth of the network grows past a certain threshold. ResNet allows for the training of deeper networks. This is enabled by skip connection identity mapping. This identity mapping does not have any parameters and is just there to add the output from the previous layer to the layer ahead. The Skip Connections between layers add the outputs from previous layers to the outputs of stacked layers ahead. This resolves the vanishing/exploding gradients problem, thereby minimizing information loss. Here, In this project Resnet34 is implemented.

ResNet34 implementation:

The residual block consisting of two convolutional layers with batchnorm between is implemented first. The 34-layer residual network is implemented by the superposition of multiple residual blocks. The residual layer is preceded by convolutional layer to normalize images for the next layer. There are 4 groups of residual layers. The first layer has 3 residual blocks with 64 input channels and 128 output channels. The second layer has 4 residual blocks with 128 input channels and 256 output channels. The

third layer has 6 residual blocks with 256 input channels and 512 output channels. The fourth layer has a residual block with 512 input channels and 512 output channels. Thus, the network has 1 initial convolutional layer and 32 convolutional layers in residual block and one last fully connected layer for output prediction, that makes for 34 layers in total.

Dataset adjustment for implementation:

The image size from mnist dataset is 28x28. This is too small for Resnet as this is smaller than the convolutional core of the first convolutional layer. Thus, the image are resized to 224x224 for Resnet and retained at 28x28 for LeNet. Also, to see the impact of grayscale, the images are also chosen from any of the three color channels.

Results:

The results show accuracy of training phase and testing phase for six different cases

a. Training Phase - LeNet - Leap dataset - Grayscale images
b. Training Phase - LeNet – Leap dataset -RGB images
c. Training Phase - LeNet – MNIST dataset
d. Test Phase - LeNet - Leap dataset - Grayscale images
e. Test Phase - LeNet – Leap dataset -RGB images
f. Test Phase - LeNet – MNIST dataset

It is observed that the LeNet performs better for MNIST dataset than for Leap dataset. In case of leap dataset, the performance of LeNet for RGB images is better than Grayscale images. This is true for both training and testing phase. The performance of ResNet is better than LeNet for both datasets. Also, the impact of selection of grayscale or RGB images is not significant. Either type of images provide similar results. The only reason to use LeNet in this scenario is for computationally low load.

An application has been developed using python to illustrate gesture recognition from images. The application automatically selects the appropriate model for recognition based on adjustable parameters.

2. Key review points

The project describes methodologies to implement a very solid goal of gesture recognition that is applied everywhere today. Two popular machine learning frameworks, The LeNet and ResNet are used to implement the same. The advantages and disadvantages of using each of the framework are clearly elucidated. Considering that most real time applications with today's technological advances uses color high resolution images, the impact of gray levels on the framework is analyzed. This is a key analysis to check if algorithms described in the past decade can be applicable today. The datasets are well chosen to be both classic benchmarks and latest technology basis for gesture recognition. The impact of gray levels is systematically studied on both frameworks. Finally, an application is reported to have been developed using python to perform gesture recognition. The application is reported to automatically select the appropriate model based on the image to perform recognition.

There are a few points of criticism those which, if implemented in this project would add significant value to the project. These points are identified below.

A key observation is that the project is missing references. For someone being introduced to LeNet or ResNet or even gesture recognition through this project, this would be a valuable source of information.

It would be easier to understand the background if references were to be included. The next point of observation is that in the description of dataset used, the MNIST dataset is said to be missing cases for j = 9 and Z = 25. The reason mentioned for this is gesture motions. This is not a very clear description of the issue. A brief note explaining the reason behind missing cases can be added. An exciting addition to the dataset was the depth map information from Kinect sensor. However, beyond the mention that this was part of the dataset, there is no mention of depth map being used. It would be exciting to understand where the depth map was used in the training process. The next comment is that the model structure of LeNet5 could be better elucidated. The functionality of each layer could be briefly discussed for ease of understanding. Additionally, in the description of LeNet architecture model, the input image is stated to be of size 28x28. This is in contradiction to the information in figure(1) where the input is stated to be 32x32. The image is 28x28 at the first convolution layer output. A minor but significant observation is the description of figure 2. Although it is evident that the project uses the ResNet34 architecture, the components of the figure 2 is not self explanatory.

3. Additional Review Points

In addition to the key points of criticism mentioned in the previous section, a few minor modifications that are good to have in the report are identified in this section. Although, these changes in no way modify the technical core contents of the project report, it makes the report more lucid. For ease of enumerating, these are listed below in a table.

Section	Point of observation	Suggested edit
Abstract	"*In order to present a better result,*"	This line seem to indicate that the results presented are being compared against some other entity. The word better indicates a comparative degree. If this is not the case, this can be modified to state, *In order to present a good result,*
1. Introduction	"*In our daily lives, we already contact with gesture recognition in a high frequency*"	The word communicate is better suited instead of contact.
	"*It can be conducted with techniques from computer vision and image processing.*"	The word implemented is better suited instead of conducted.
3.2. ResNet34	"*But when the network goes deeper, the result and may*	The sentence has been complicated to the point that the meaning is not conveyed. This can be summarized along

	not be so well because a lot of problem raising by the depth of the network."	the lines of *Deeper networks have problems due to the depth.*
4. Implementation	*seperatet*	Spell Check recommended
	"which corresponding to implement the models, resizing of the dataset, and the whole training process"	The better grammatical representation would be *"which correspond to implementation of the models, resizing of the dataset, and the training process"*
4.1 Model Implementation LeNet5	*"And of course, a smaller size of the model always corresponding to a relatively poorer performance"*	*"..Corresponds to.."*
	"There is only 7 layers"	*"There are only 7 layers"*
	"LeNet is consist of"	*"LeNet consists of"*
	"…we need to make the LeNet5 enable to deal…"	*"…we need to make the LeNet5 be able to deal…"*
ResNet34	*"Since Resnet34 got a much.."*	*"Since Resnet34 has a much.."*
	"…we first have a convolutional layer for the normalize of the input images…"	*"…we first have a convolutional layer for the normalization of the input images…"*
	blcok	Spell check recommended
	…"network end up with a.."	*…"network ends up with a.."*
4.2 Dataset Adjustment	*"…This makes us a problem that.."*	*"…This poses a problem that.."*
	Layre	Spell check recommended
	"…LeNet5 can neither learning an image with too large scale."	The meaning of this sentence is not clear.
4.3 Training Process	Title	Spell check recommended
	"For thr criterion.."	Spell check recommended
5. Test Result	*"…how the loss variate…"*	*"…how the loss varies…"*
	"…clearly slowdown.."	*"…clear slowdown."*
	Basicly	Spell Check recommended
	Raletively	Spell Check recommended
	Affect	*Effect*
	Mentains	*maintains*
	"…Since every gray scale image is actually generate from.."	*generated*
	Gray sacle images	Spell Check recommended

6.Application	*"….To make the user more convenient.."*	The implication here is to make it convenient for the user.
	"…prediction work by analyze.."	"analyzing"

4. Review Summary

Overall, the problem statement of the project is a classic one and appeals to stalwarts and newbies alike. With the addition of a few features like references, illustration of the use of depth map in the dataset and more detailed explanation of the LeNet architecture levels, the project would cater to wide category of technologists. Additionally, the rectification of minor grammatical glitches will make for a better read.