"TRAFFIC SIGN RECOGNITION FOR AUTONOMOUS SYSTEMS USING DEEP LEARNING"

A Report

Submitted as special assignment

of

2ICDE61 DEEP LEARNING FOR VISION SYSTEM

By
Deep Khut (19BIC008)

Under the Guidance of
Prof. Harsh Kapadia

INSTRUMENTATION AND CONTROL ENGINEERING
INSTITUTE OF TECHNOLOGY
NIRMA UNIVERSITY
Ahmedabad 382 481

NOVEMBER 2022

INTRODUCTION

TRAFFIC SIGNS ARE AN INTEGRAL PART OF OUR ROAD INFRASTRUCTURE. THEY PROVIDE CRITICAL INFORMATION, SOMETIMES COMPELLING RECOMMENDATIONS, FOR ROAD USERS, WHICH IN TURN REQUIRES THEM TO ADJUST THEIR DRIVING BEHAVIOR TO MAKE SURE THEY ADHERE WITH WHATEVER ROAD REGULATION CURRENTLY ENFORCED. WITHOUT SUCH USEFUL SIGNS, WE WOULD MOST LIKELY BE FACED WITH MORE ACCIDENTS, AS DRIVERS WOULD NOT BE GIVEN CRITICAL FEEDBACK ON HOW FAST THEY COULD SAFELY GO, OR INFORMED ABOUT ROAD WORKS, SHARP TURN, OR SCHOOL CROSSINGS AHEAD. IN OUR MODERN AGE, AROUND 1.3M PEOPLE DIE ON ROADS EACH YEAR. THIS NUMBER WOULD BE MUCH HIGHER WITHOUT OUR ROAD SIGNS. NATURALLY, AUTONOMOUS VEHICLES MUST ALSO ABIDE BY ROAD LEGISLATION AND THEREFORE RECOGNIZE AND UNDERSTAND TRAFFIC SIGNS.

TRADITIONALLY, STANDARD COMPUTER VISION METHODS WERE EMPLOYED TO DETECT AND CLASSIFY TRAFFIC SIGNS, BUT THESE REQUIRED CONSIDERABLE AND TIME-CONSUMING MANUAL WORK TO HANDCRAFT IMPORTANT FEATURES IN IMAGES. INSTEAD, BY APPLYING DEEP LEARNING TO THIS PROBLEM, WE CREATE A MODEL THAT RELIABLY CLASSIFIES TRAFFIC SIGNS, LEARNING TO IDENTIFY THE MOST APPROPRIATE FEATURES FOR THIS PROBLEM BY ITSELF.

DATASET DESCRIPTION

THE IMAGE DATASET IS CONSISTS OF MORE THAN 50,000 PICTURES OF VARIOUS TRAFFIC SIGNS(SPEED LIMIT, CROSSING, TRAFFIC SIGNALS, ETC.) AROUND 43 DIFFERENT CLASSES ARE PRESENT IN THE DATASET FOR IMAGE CLASSIFICATION. THE DATASET CLASSES VARY IN SIZE LIKE SOME CLASS HAS VERY FEW IMAGES WHILE OTHERS HAVE A VAST NUMBER OF IMAGES. THE DATASET DOESN'T TAKE MUCH TIME AND SPACE TO DOWNLOAD AS THE FILE SIZE IS AROUND 314.36 MB. IT CONTAINS TWO SEPARATE FOLDERS, TRAIN AND TEST, WHERE THE TRAIN FOLDER IS CONSISTS OF CLASSES, AND EVERY CATEGORY CONTAINS VARIOUS IMAGES.

THE DATASET IS SPLIT INTO TRAINING, TEST AND VALIDATION SETS, WITH THE FOLLOWING CHARACTERISTICS:

- IMAGES ARE 32 (WIDTH) X 32 (HEIGHT) X 3 (RGB COLOR CHANNELS)
- TRAINING SET IS COMPOSED OF 34799 IMAGES
- VALIDATION SET IS COMPOSED OF 4410 IMAGES
- TEST SET IS COMPOSED OF 12630 IMAGES
- THERE ARE 43 CLASSES (E.G. SPEED LIMIT 20KM/H, NO ENTRY, BUMPY ROAD, ETC.)

0 . Class : Speed limit (20km/h)

1 . Class : Speed limit (30km/h)

2 . Class : Speed limit (50km/h)

3 . Class : Speed limit (60km/h)

4 . Class : Speed limit (70km/h)

5 . Class : Speed limit (80km/h)

Proposed Methods

- AlexNet

The original AlexNet architecture was proposed for the ImageNet data which is much larger then the images we have in the GTSRB data set. I therefore build a much smaller CNN then the original architecture proposed in the AlexNet paper, but build it according to the same design principles. The only difference being the local response normalization layers used in the originally proposed AlexNet model are not included as these have fell out of favor in recent times. I instead include batch normalization layers, this essentially incorporates normalization within each layer of the network and allows the network to reduce the internal co-variate shift via learnt parameters. The result of this is an increase in training speed and an increased robustness to choices in weight initialization.

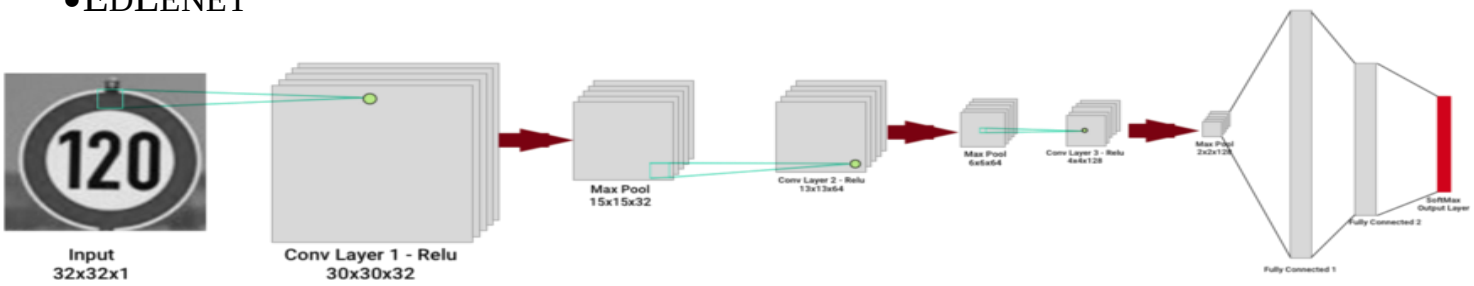
Layer	Description	Input	Output
Convolution 5x5	1x1 stride, Same padding	**32x32x3**	32x32x64
Batch Normalisation	Decay: 0.999, eps: 0.001	32x32x64	32x32x64
ReLU Activation		32x32x64	32x32x64
Max pooling	2x2 stride, 3x3 window	32x32x64	16x16x64
Convolution 5x5	1x1 stride, Same padding	16x16x64	16x16x64
Batch Normalisation	Decay: 0.999, eps: 0.001	16x16x64	16x16x64
ReLU Activation		16x16x64	16x16x64
Max pooling	2x2 stride, 3x3 window	16x16x64	8x8x64
Flatten	3 dimensions -> 1 dimension	8x8x64	4096
Fully Connected	connect every neuron from layer above	4096	384
Batch Normalisation	Decay: 0.999, eps: 0.001	384	384
ReLU Activation		384	384
Dropout	Keep Prob: 0.8	384	384
Fully Connected	connect every neuron from layer above	384	192
Batch Normalisation	Decay: 0.999, eps: 0.001	192	192
ReLU Activation		192	192
Dropout	Keep Prob: 0.8	192	192
Fully Connected	output = number of traffic signs in data set	192	**43**

- DENSENET-121

IN DENSENET EACH LAYER IS CONNECTED TO EVERY OTHER LAYER IN THE NETWORK IN A FEED FORWARD FASHION. FOR EACH LAYER, THE FEATURE-MAPS OF ALL PRECEDING LAYERS ARE USED AS INPUT, AND ITS OWN FEATURE-MAPS ARE CONCATENATED WITH ITS INPUT INTO A SINGLE TENSOR AND THE USED AS INPUTS INTO ITS SUBSEQUENT LAYER. A STANDARD FEED FORWARD CNN WITH L LAYERS WILL HAVE L CONNECTIONS (ONE BETWEEN EACH LAYER), DENSENET WITH ITS DENSELY CONNECTED SCHEME MUST HAVE (L+1)/ 2 DIRECT CONNECTIONS.

LAYERS	**OUTPUT SIZE**	**DENSENET-121**
CONVOLUTION	112X112	7 X 7 CONV, STRIDE2
POOLING	56X56	3 X 3 MAX POOLING, STRIDE2
DENSE BLOCK (1)	56 X 56	$\begin{bmatrix} \mathbf{1} & \boldsymbol{X} & \mathbf{1}\,\boldsymbol{conv} \\ \mathbf{3} & \boldsymbol{X} & \mathbf{3}\,\boldsymbol{conv} \end{bmatrix}$ **X 6**
TRANSITION LAYER (1)	56 X 56	1 X 1 CONV
	28 X 28	2 X 2 AVERAGE POOL, STRIDE 2
DENSE BLOCK (2)	28 X 28	$\begin{bmatrix} \mathbf{1} & \boldsymbol{X} & \mathbf{1}\,\boldsymbol{conv} \\ \mathbf{3} & \boldsymbol{X} & \mathbf{3}\,\boldsymbol{conv} \end{bmatrix}$ **X 6**
TRANSITION LAYER (2)	28 X 28	1 X 1 CONV
	14 X 14	2 X 2 AVERAGE POOL, STRIDE 2
DENSE BLOCK (3)	14 X 14	$\begin{bmatrix} \mathbf{1} & \boldsymbol{X} & \mathbf{1}\,\boldsymbol{conv} \\ \mathbf{3} & \boldsymbol{X} & \mathbf{3}\,\boldsymbol{conv} \end{bmatrix}$ **X 6**
TRANSITION LAYER (2)	14 X 14	1 X 1 CONV
	7 X 7	2 X 2 AVERAGE POOL, STRIDE 2
DENSE BLOCK (4)	7 X 7	$\begin{bmatrix} \mathbf{1} & \boldsymbol{X} & \mathbf{1}\,\boldsymbol{conv} \\ \mathbf{3} & \boldsymbol{X} & \mathbf{3}\,\boldsymbol{conv} \end{bmatrix}$ **X 6**
CLASSIFICATION LAYER	1 X 1	7 X 7 GLOBAL AVERAGE
		1000D FULLY CONNECTED, SOFTMAX

- EDLENET



THE NETWORK IS COMPOSED OF 3 CONVOLUTIONAL LAYERS — KERNEL SIZE IS 3X3, WITH DEPTH DOUBLING AT NEXT LAYER — USING RELU AS THE ACTIVATION FUNCTION, EACH FOLLOWED BY A 2X2 MAX POOLING OPERATION. THE LAST 3 LAYERS ARE FULLY CONNECTED, WITH THE FINAL LAYER PRODUCING 43 RESULTS (THE TOTAL NUMBER OF POSSIBLE LABELS) COMPUTED USING THE SOFTMAX ACTIVATION FUNCTION. THE NETWORK IS TRAINED USING MINI-BATCH STOCHASTIC GRADIENT DESCENT WITH THE ADAM OPTIMIZER.

- **CUSTOMIZED VERSION**

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 28, 28, 32)        2432
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 24, 24, 32)        25632
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 12, 12, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 12, 12, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 10, 10, 64)        18496
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 8, 8, 64)          36928
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 4, 4, 64)          0
_________________________________________________________________
dropout_1 (Dropout)          (None, 4, 4, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 1024)              0
_________________________________________________________________
dense (Dense)                (None, 256)               262400
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 43)                11051
=================================================================
Total params: 356,939
Trainable params: 356,939
Non-trainable params: 0
_________________________________________________________________
```

RESULTS

- ## ALEXNET

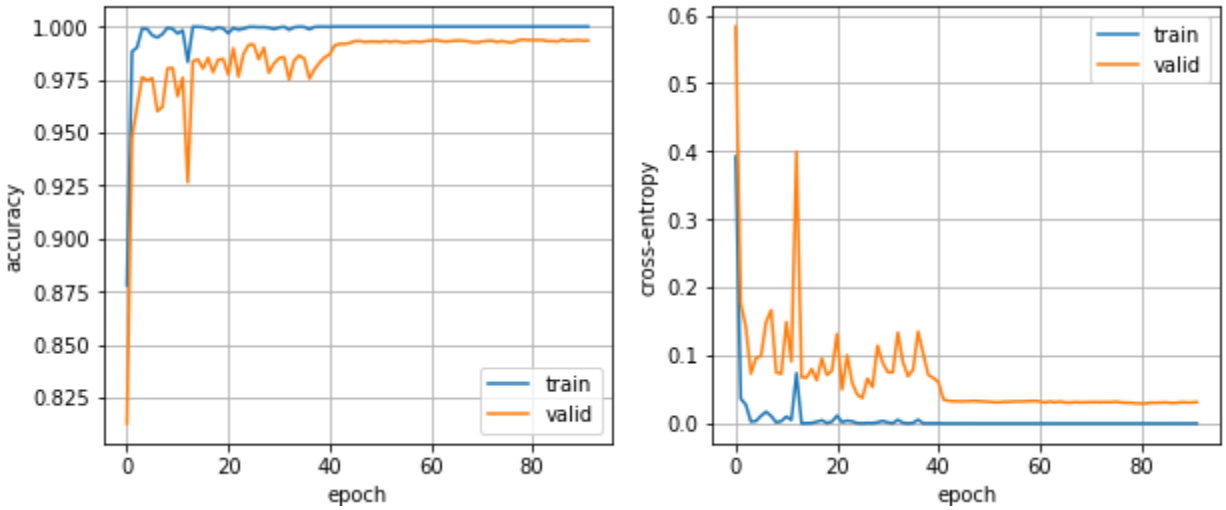


FIG1. ACCURACY AND LOSS FOR ALEXNET

- ## DENSENET

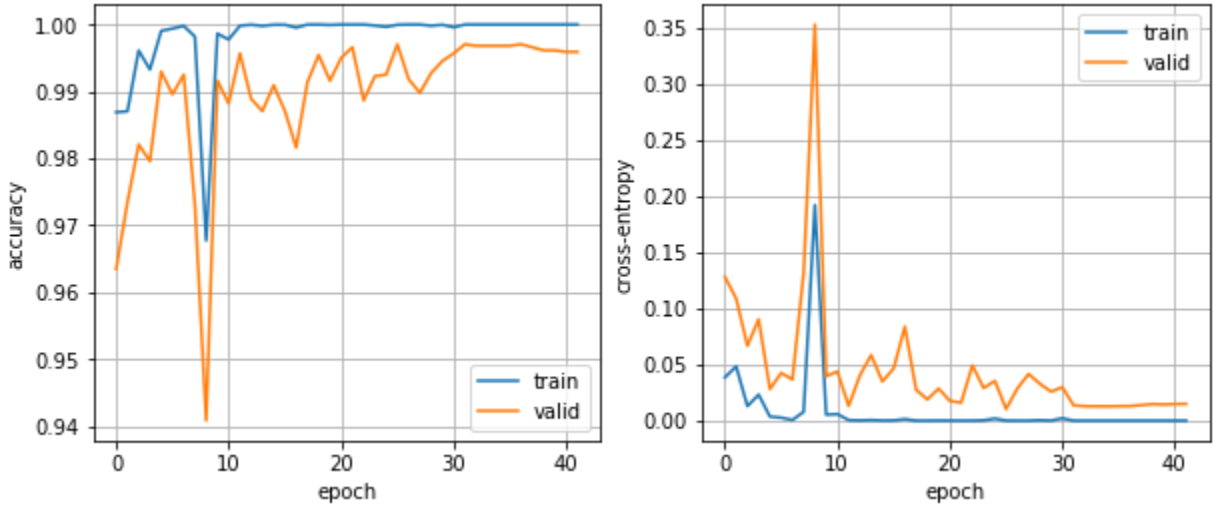


FIG2. ACCURACY AND LOSS FOR DENSENET

- **EDLENET**

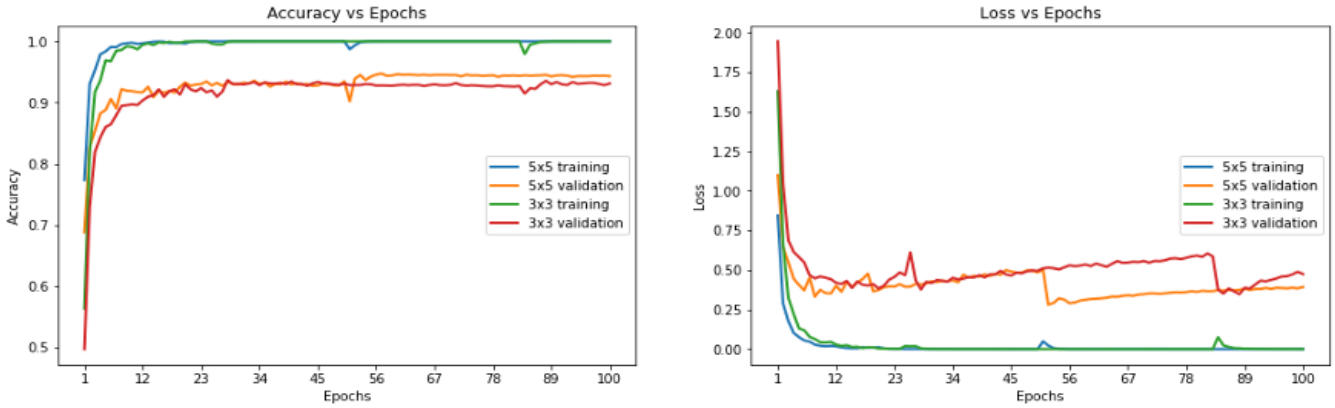


FIG3. ACCURACY AND LOSS FOR EDLENET

- **CUSTOMIZED VERSION**

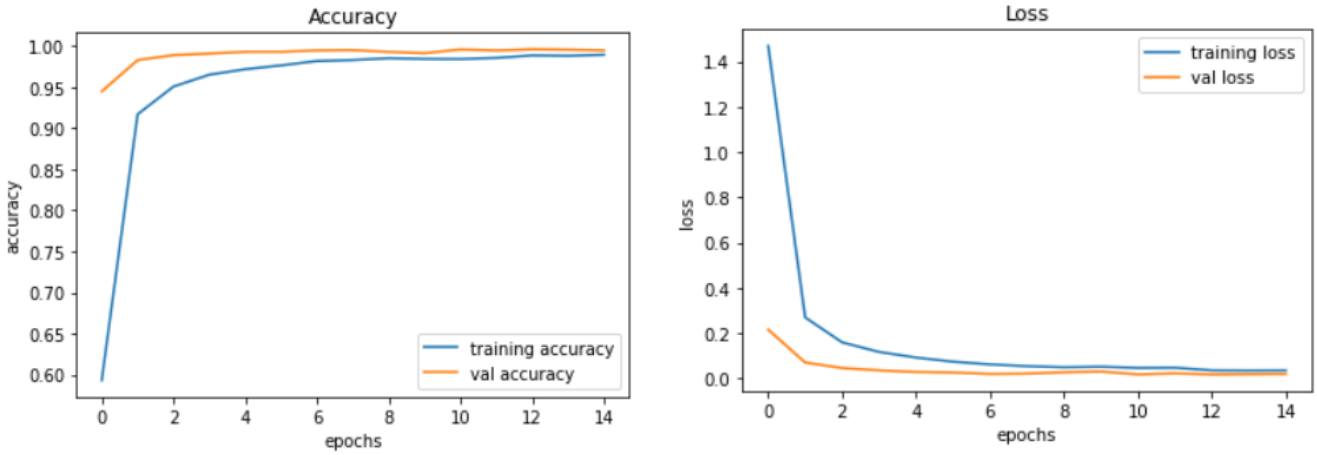


FIG4. ACCURACY AND LOSS

TRANSFER LEARNING MODEL	VALIDATION ACCURACY
ALEXNET	TRAINING ACCURACY: 100 % VALIDATION ACCURACY: 99.8% TEST ACCURACY: 98.32 %
DENSENET	TRAINING ACCURACY: 100% VALIDATION ACCURACY: 99.7% TEST ACCURACY 99.02%
EDLENET	TRAINING ACCURACY: 99.8% VALIDATION ACCURACY: 99.02% TEST ACCURACY 97.86%
CUSTOMIZED	TRAINING ACCURACY: 99.6% VALIDATION ACCURACY: 98.4% TEST ACCURACY 97.32%

CONCLUSION & FUTURE SCOPE OF WORK

THROUGH THIS PROJECT, UNDERSTOOD HOW DEEP LEARNING CAN BE USED TO CLASSIFY TRAFFIC SIGNS WITH HIGH ACCURACY, EMPLOYING A VARIETY OF PRE-PROCESSING AND REGULARIZATION TECHNIQUES, AND TRYING DIFFERENT MODEL ARCHITECTURES.

ALEXNET AND DENSENET HAS PERFORMED VERY GOOD IN THIS PROBLEM GIVING AN ACCURACY OF 98 AND 99% RESPECTIVELY. WITH THE ACCURACY AND REAL-TIME RESULTS, THE PROPOSED MODEL CAN BE IMPLEMENTED/USED IN AUTONOMOUS AS WELL AS NON-AUTONOMOUS VEHICLES WHERE IT CAN BE USED TO ALERT THE DRIVER REGARDING THE ROAD SIDE TRAFFIC SIGNS THAT ARE COMING UP OR PASSING BY.