Asif Zubayer Palak [20101179]
Israt Jahan Tonni [19101591]
Masuda Rahman Fatima [20101015]
Maruf Bin Murtuza [20101334]

CSE422 LAB REPORT

SOLVING CAPTCHA BY IMAGE PROCESSING

INTRODUCTION

SOLVING CAPTCHA BY IMAGE PROCESSING

THE COMPLETELY AUTOMATED PUBLIC TURING TEST TO TELL COMPUTERS AND HUMANS APART (CAPTCHA) IS A CRUCIAL HUMAN-MACHINE DISTINGUISHING TECHNIQUE FOR WEBSITES TO FEND OFF AUTOMATIC ATTACKS FROM MALICIOUS SOFTWARE. STUDIES ON CAPTCHA RECOGNITION HAVE THE POTENTIAL TO UNCOVER SECURITY FLAWS, ADVANCE CAPTCHA TECHNOLOGY, AND EVEN ADVANCE HANDWRITING AND LICENSE PLATE RECOGNITION. IN ORDER TO RECOGNIZE CAPTCHA AND AVOID USING CONVENTIONAL IMAGE PROCESSING TECHNIQUES LIKE LOCATION AND SEGMENTATION, THIS RESEARCH SUGGESTED A SOLUTION BASED ON THE CONVOLUTIONAL NEURAL NETWORK (CNN) MODEL. THE PROBLEM OF OVERFITTING AND LOCAL OPTIMAL SOLUTION HAS BEEN RESOLVED, AND THE ADAPTIVE LEARNING RATE IS INTRODUCED TO QUICKEN THE MODEL'S CONVERGENCE. USING THE MULTI TASK JOINT TRAINING MODEL, MODEL RECOGNITION ACCURACY AND GENERALIZABILITY ARE ENHANCED.

METHODOLOGY

I. DATASET DESCRIPTION

SINCE OUR CAPTCHA BREAKING PROBLEM IS A MACHINE LEARNING PROBLEM, WE NEED A HUGE AMOUNT OF DATA TO TRAIN OUR MODELS. FOR OUR PROJECT, WE USED A DATASET FROM KAGGLE THAT HAD AROUND 9000 SCRAPED IMAGES FROM SIMPLE CAPTCHA. THE DATASET CONSISTS OF 4 CHARACTERS ALPHANUMERIC DATA AS MENTIONED IN THEIR FILE NAME. EACH IMAGE IS OF SIZE 72*24 AND IS IN BLACK AND WHITE. TOTAL NUMBER OF IMAGES IN THE DATASET IS 9955 FROM WHICH 7467 SAMPLES WERE SELECTED FOR TRAINING AND THE REST FOR TESTING OR VALIDATION. WE USED GOOGLE DRIVE TO STORE THE DATASET AND GOOGLE COLAB FOR OUR ENTIRE PROJECT. THE IMAGES WERE READ AS GRAYSCALE USING OPENCV WITH THEIR RESPECTIVE INPUT SIZES. BELOW ARE A COUPLE OF SAMPLE IMAGES FROM OUR DATASET.

Examples from the dataset

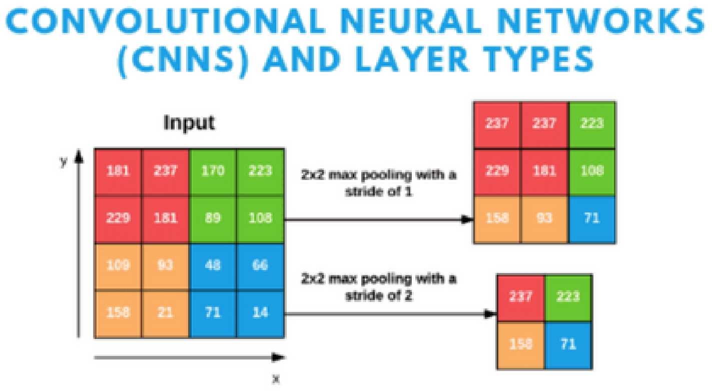
II. PRE-PROCESSING

SOME OF THE ADVANTAGES THAT CAME FROM THE DATASET DOWNLOADED FROM KAGGLE WAS THAT, VERY LITTLE PREPROCESSING WAS REQUIRED FOR FILE NAMING AND CONSISTENCY. ALL THE FILES NAMES ARE LABELED ACCORDING TO THE CAPTCHAS. WE MADE SURE THAT THE DATASET IS UNIQUE DURING THE AUTO-GENERATION PROCESS. BEFORE BEING TRANSFERRED, THE DATASET IS USUALLY PREPROCESSED BEFORE BEING TRANSMITTED. OUR PREPROCESSING CONSISTS OF MAKING IT HAVE ONLY 1 COLOR CHANNEL (BLACK AND WHITE). THE CAPTCHA IS NO MORE THAN 4 LETTERS. IN THIS PROCESS, THE IMAGES ARE BEING SCALED BETWEEN 0 AND 1 FROM 0 TO 255. IT ALSO RESHAPES THE IMAGE TO WIDTH 72, HEIGHT 24, CHANNEL 1. THE IMAGES WERE READ AS GRAYSCALE USING OPENCV WITH THEIR RESPECTIVE INPUT SIZES. THIS HELPS US TO REMOVE THE NOISE TO SOME EXTENT AS THE IMAGES GET INVARIANT OF THE BACKGROUND COLOR. EACH GRAYSCALE IMAGE IS THEN SCALED AND RESHAPED. SIMULTANEOUSLY, THE FILE NAMES ARE ALSO BEING CONVERTED INTO NUMPY ARRAYS OF DIMENSION 4*34, WHERE 34 REPRESENTS ALL THE DIFFERENT CHARACTERS FOUND IN THE LABELS AND 4 FOR EACH CHARACTER IN CAPTCHA. ACCORDING TO THIS LIST OF CHARACTERS, WE UPDATE THE APPROPRIATE LOCATION TO ONE IN THIS ARRAY AT EACH PLACE FOR EACH CHARACTER. AS A RESULT, FOLLOWING THIS PRE-PROCESSING, WE HAVE ALL THE GRAYSCALE IMAGES AS NUMPY ARRAYS AS WELL AS THE TARGET ARRAY THAT CONTAINS DATA ON THE CHARACTERS FOUND IN EACH CAPTCHA IMAGE.

III. MODELS APPLIED

1.CONVOLUTIONAL NEURAL NETWORK (CNN)

IN 1995, YANN LECUN ET AL. [1] INTRODUCED THE CONCEPT OF CONVOLUTIONAL NEURAL NETWORKS IN WHICH THEY TRIED TO RECOGNIZE HANDWRITTEN CHARACTERS. A CONVOLUTIONAL NEURAL NETWORK (CNN) IS A VARIANT OF MULTILAYER PERCEPTRON. CNN CONTAINS MANY LAYERS, OF WHICH SOME COULD BE CONVOLUTIONAL LAYERS. A CONVOLUTIONAL LAYER IS A LAYER THAT APPLIES A CONVOLUTION FILTER (A GREAT WAY TO PROCESS IMAGES FOR CERTAIN FEATURES). IT CAN BE SEEN AS A SLIDING WINDOW FUNCTION APPLIED TO INPUT PIXEL MATRIX. FOR THIS PROJECT WE USED CONVOLUTIONAL NEURAL NETWORK (CNN) AS OUR FIRST MODEL FOR DETECTING THE CAPTCHA PRESENT IN THE IMAGE. FOR THIS PURPOSE, THE TRAINING SAMPLE DATASET IS FIRST PRE-PROCESSED AND THEN THE MODEL IS DEVELOPED CONSISTING OF 24 LAYERS. THE TOTAL NUMBERS OF PARAMETERS ARE 1,818,196 WHERE 1,818,132 PARAMETERS ARE TRAINABLE AND 64 ARE NON-TRAINABLE PARAMETERS. A BRIEF ARCHITECTURE OF THE LAYERS IS DEPICTED IN FIGURE.
THE FIRST LAYER IS THE INPUT LAYER WHICH TAKES THE IMAGE AS INPUT. THEN WE HAVE CONVOLUTIONAL LAYERS AND THE MAX POOLING LAYERS WHICH EXTRACTS THE MOST PROMINENT FEATURES FROM THE IMAGES. THE NEXT LAYER IS THE BATCH NORMALIZATION LAYER, USED TO IMPROVE THE STABILITY OF THE MODEL. THEN WE HAVE A FLATTEN LAYER WHICH CONVERTS THE INPUT FROM MAX POOL LAYER TO A LONG VECTOR OF DESIRED DIMENSIONS. THIS IS DONE SO AS THE FURTHER NEURAL NETWORK IS EASILY PROCESSED AND THE BACK PROPAGATION IS CARRIED OUT EASILY. FURTHER THERE ARE FIVE DENSE LAYERS IN THIS NEURAL NETWORK EACH OF WHICH IS CONNECTED TO THE FLATTEN LAYER. EACH OF THESE DENSE LAYERS DEPLOYS THE ACTIVATION FUNCTION 'RELU' TO TRAIN THE PARAMETERS.


CONVOLUTIONAL NEURAL NETWORKS (CNNS) AND LAYER TYPES
Input
y
x
181 237 170 223
229 181 89 108
109 93 48 66
158 21 71 14
2x2 max pooling with a stride of 1
237 237 223
229 181 108
158 93 71
2x2 max pooling with a stride of 2
237 223
158 71

CNN layer types

FURTHER TO EACH OF THESE DENSE LAYERS IS CONNECTED A DROPOUT LAYER USED FOR REGULARIZATION. FOLLOWING THE DROPOUT LAYER IS AGAIN A DENSE LAYER WHICH USES THE ACTIVATION FUNCTION 'SIGMOID'. THE SIGMOID FUNCTION IS ALSO CALLED A LOGISTIC FUNCTION AND IT TRANSFORMS THE INPUT TO VALUES BETWEEN 0 AND 1.

AFTER FORMING THE MODEL, THE LOSS FUNCTION IS CALCULATED USING THE CATAGORICAL CROSS ENTROPY AND ADAM IS USED AS THE OPTIMISER WITH A LEARNING RATE OF 0.001

2. RECURRENT NEURAL NETWORK (RNN)

IN RNNS, CONNECTIONS BETWEEN UNITS HAVE A DIRECTED CYCLE. RNNS ARE CALLED RECURRENT BECAUSE THEY PERFORM THE SAME TASK FOR EACH ELEMENT IN A SEQUENCE WITH THE OUTPUT BEING DEPENDENT ON THE PREVIOUS COMPUTATIONS, UNLIKE FEED-FORWARD NETWORKS. RNNS MAKE USE OF THEIR INTERNAL MEMORY TO PROCESS ARBITRARY INPUT SEQUENCES. ONE OF THE COMMON APPLICATIONS OF RNNS INCLUDE HANDWRITING RECOGNITION.

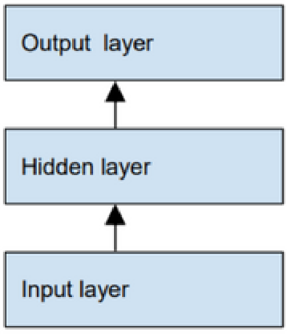


Fig 2: A Sample feed-forward network

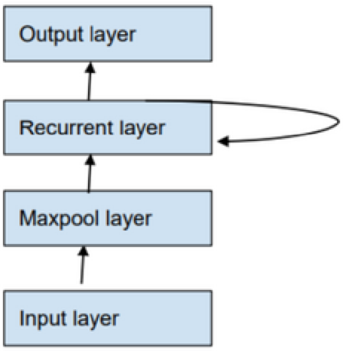
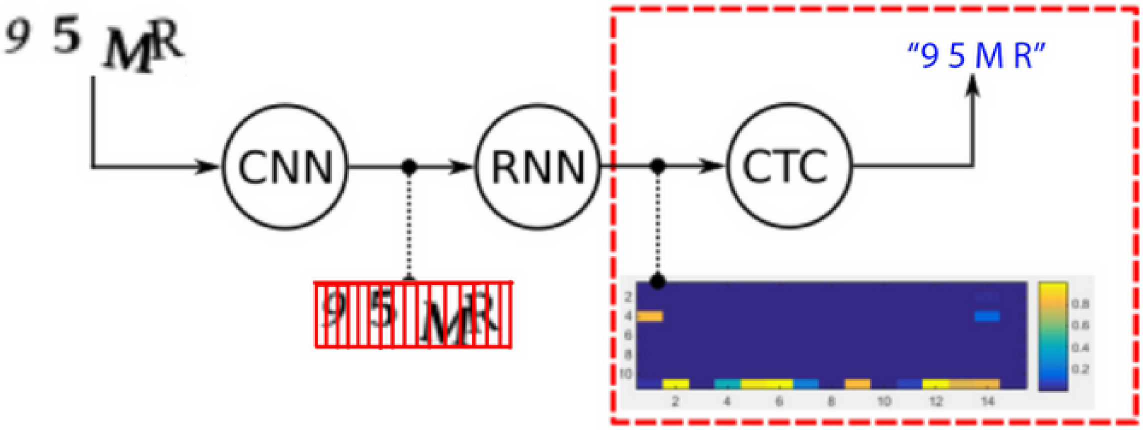


Fig 3: An example of RNN.

ONE OF THE MOST COMMONLY USED RNN IS LSTM (LONG SHORT-TERM MEMORY), WHICH WE HAVE USED IN OUR PROJECT.
IN A SEPERATE PRE PROCESSING, WE SLICED THE IMAGE IN 8 PIECES INSTEAD OF USING BOUNDING BOXES AND CHOOSE 8 FOR OUR TIME STEP IN THE LSTM LAYER (RNN). THIS SOLVES THE PROBLEM THAT COMES WITH USING BOUNDING BOXES WHEN DEALING WITH OVERLAPPING CHARACTERS. USING THIS RNN, AS LONG AS EACH CHARACTER IS SEPERATED BY TWO OR THREE PARTS TO BE PROCESSED AND DECODED LATER THEN THE SPACING BETWEEN EACH CHARACTER IS IRRELEVANT.
AFTER STACKED LSTM LAYERS WITH SOFTMAX (SM) ACTIVATION FUNCTION, WE HAVE CTC TO OPTIMIZE OUR PROBABILITY TABLE.



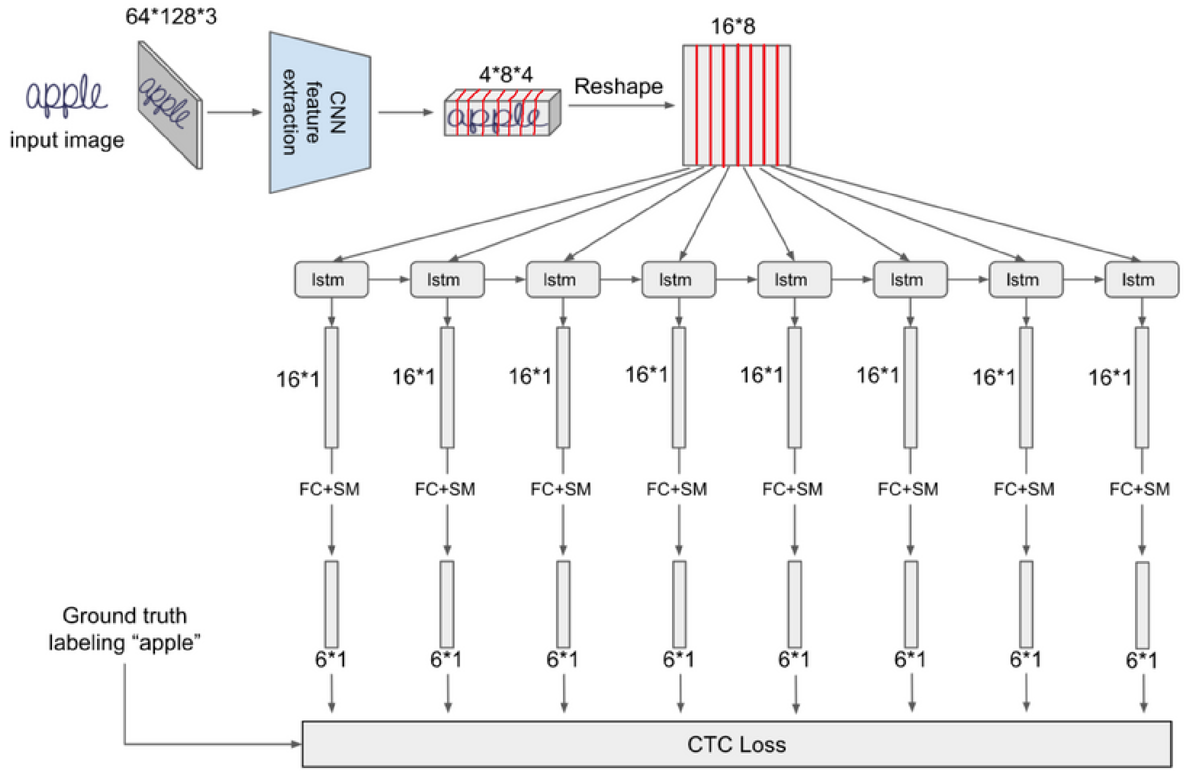
3. CONNECTIONIST TEMPORAL CLASSIFICATION LOSS (CTC)

IN ORDER TO TRAIN RECURRENT NEURAL NETWORKS LIKE LSTM NETWORKS TO HANDLE SEQUENCE PROBLEMS WHERE THE TIME IS UNCERTAIN, CONNECTIONIST TEMPORAL CLASSIFICATION (CTC), IS USED, WHICH IS A TYPE OF NEURAL NETWORK OUTPUT. IT CAN BE APPLIED TO TASKS LIKE PHONE RECOGNITION IN SPOKEN AUDIO OR ONLINE HANDWRITING RECOGNITION. CTC RELATES TO THE RESULTS AND SCORING AND HAS NOTHING TO DO WITH THE FUNDAMENTAL DESIGN OF THE NEURAL NETWORK.

WITHOUT FIRST SEGMENTING THE INPUT, CONNECTIONIST TEMPORAL CLASSIFICATION—CTC—CAN ALIGN VARIABLE-LENGTH INPUT SEQUENCES TO VARIABLE-LENGTH TARGETS. END-TO-END UN-SEGMENTED SEQUENCE LABELING SYSTEMS THAT INCORPORATE HANDWRITING RECOGNITION CAN BE ACCOMPLISHED WITH THE AID OF CTC.

THE INPUT CONSISTS OF A SERIES OF OBSERVATIONS, WHILE THE OUTPUTS—WHICH MAY ALSO INCLUDE BLANK OUTPUTS—CONSIST OF A SERIES OF LABELS. BECAUSE THERE ARE FAR MORE OBSERVATIONS THAN LABELS, TRAINING IS CHALLENGING. THE CONTINUOUS OUTPUT OF A CTC NETWORK IS FITTED DURING TRAINING TO REPRESENT THE LIKELIHOOD OF A LABEL.

CTC DOES NOT TRY TO LEARN TIMINGS AND BOUNDARIES: IF THE ONLY DIFFERENCE BETWEEN TWO LABEL SEQUENCES IS HOW THEY ALIGN, BLANKS ARE NOT TAKEN INTO ACCOUNT.

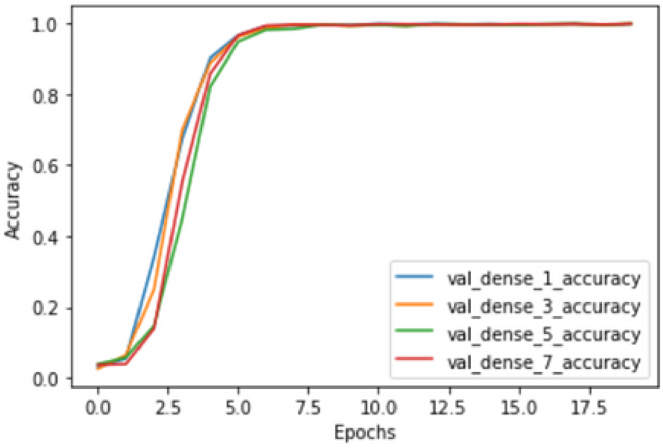


Model summery

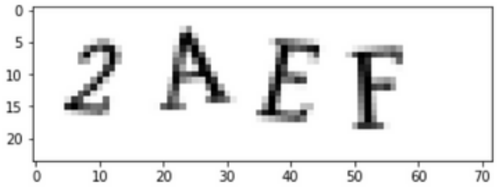
IV. RESULTS

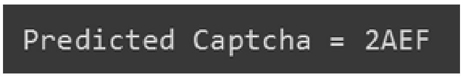
RESULTS AFTER USING THE CNN MODEL WITH 24 LAYERS:

20 EPOCHS WERE USED TO GET AN AVERAGE ACCURACY OF 99.73% FOR EACH OF THE 4 CHARACTERS IN THE CAPTCHA.



PREDICTION ON A SAMPLE IMAGE:

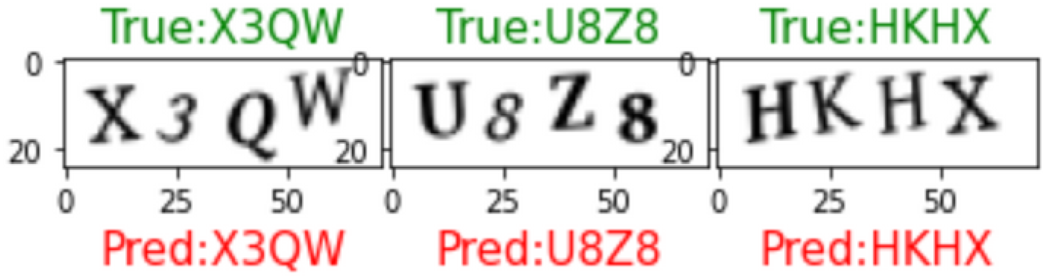




```
Predicted Captcha = 2AEF
```

RESULTS AFTER USING THE COMBINED CNN, RNN, CTC MODELS:

AFTER USING 10 EPOCHS, THE ACCURACY GOT TO 96% AND LOSS WAS AT ABOUT 0.54
HERE ARE SOME PREDICTIONS ON SAMPLE CAPTCHA:



REFERENCE

[1] Y.LECUN, B. BOSER, J. S. DENKE, D. HENDERSON, R. E. HOWARD, W. HUBBARD AND L.D. JACKEL. HANDWRITTEN DIGIT RECOGNITION WITH A BACK-PROPAGATION NETWORK. 1989.

[2] ORIOL VINYALS, ALEXANDER TOSHEV, SAMY BENGIO, DUMITRU ERHAN, SHOW AND TELL: A NEURAL IMAGE CAPTION GENERATOR, 20 APR 2015.

[3] H. JAEGER. HARNESSING NONLINEARITY: PREDICTING CHAOTIC SYSTEMS AND SAVING ENERGY IN WIRELESS COMMUNICATION. SCIENCE, 2004.

[4] W. MAASS, T. NATSCHLÄGER, AND H. MARKRAM. A FRESH LOOK AT REAL-TIME COMPUTATION IN GENERIC RECURRENT NEURAL CIRCUITS. TECHNICAL REPORT, INSTITUTE FOR THEORETICAL COMPUTER SCIENCE, TU GRAZ, 2002.