A

Project Report

On

Traffic Sign Detection

(CE255 – Software Group Project)

Prepared by

20CE126 – Hardi Shah

20CE130 – Krima Shah

20CE133 – Prachi Shah

Under the Supervision of

Mr. Deepkumar Kothadiya

Submitted to

Charotar University of Science & Technology (CHARUSAT)

for the Partial Fulfillment of the Requirements for the

Degree of Bachelor of Technology (B.Tech.)

in Computer Engineering (CE)

for 4th semester B.Tech.

Submitted at

DEPARTMENT OF COMPUTER ENGINEERING

Chandubhai S. Patel Institute of Technology (CSPIT),CHARUSAT

At: Changa, Dist: Anand, Pin: 388421.

April 2022

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

Accredited with Grade A by NAAC

Accredited with Grade A by KCG

cspit
Chandubhai S Patel Institute of Technology

This is to certify that the report entitled "**Traffic Sign Detection**" is a bonafied work carried out by **Hardi Shah(20CE126), Krima Shah(20CE130), Prachi Shah(20CE133)** under the guidance and supervision of Mr. Deepkumar Kothadiya for the subject **Software Group Project (CE255)** of 4th Semester of Bachelor of Technology in **Computer Engineering** at Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology & Engineering (FTE) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred by the examiner(s).

Under the Supervision of,

Mr. Deepkumar Kothadiya,
Dept. of Computer Engineering,
CSPIT, CHARUSAT, Changa, Gujarat.

Dr. Ritesh Patel,
Head – Department of Computer Engineering,
CHARUSAT, Changa, Gujarat.

Table of Contents

1. Introduction

Advanced driver assistance systems (ADAS) are one of the fastest-growing fields in automotive electronics. ADAS technology can be based upon vision systems, active sensors technology, car data networks,etc. These devices can be utilized to extract various kinds of data from the driving environments

One of the most important difficulties that ADAS face is the understanding of the environment and guidance of the vehicles in real outdoor scenes. Traffic signs are installed to guide, warn, and regulate traffic. They supply information to help drivers.

In current traffic management systems, there is a high probability that the driver may miss some of the traffic signs on the road because of overcrowding due to neighbouring vehicles. With the continuous growth of vehicle numbers in urban agglomerations around the world, this problem is only expected to grow worse.

In the real world, drivers may not always notice road signs. At night or in bad weather, traffic signs are harder to recognize correctly and the drivers are easily affected by headlights of oncoming vehicles. These situations may lead to traffic accidents and serious injuries.

A vision-based road sign detection and recognition system is thus desirable to catch the attention of a driver to avoid traffic hazards. These systems are important tasks not only for ADAS, but also for other real-world applications including urban scene understanding, automated driving, or even sign monitoring for maintenance.It can enhance safety by informing the drivers about the current state of traffic signs on the road and giving valuable information about precaution.

Traffic-sign detection is a technology by which an automated vehicle is able to detect the traffic signs put on the road. For example, speed limit, Bump ahead, Railway Crossing, Oneway Traffic, etc .

It uses image processing techniques to detect the traffic signs. The detection methods can be generally divided into color based, shape based and learning based methods. Traffic signs can be analyzed using forward-facing cameras in many modern cars, vehicles and trucks. One of the basic use cases of a traffic-sign detection system is for speed limits.

Modern traffic-sign detection systems are being developed using Machine Learning, Deep Learning, Vision transformers, etc.

A visual-based traffic sign detection system can be implemented on the an automobile with an aim of detecting and recognizing all emerging traffic signs. The same would be displayed to the driver with alarm-triggering features if the driver refuses to follow the traffic signs.

We can used Traffic Sign detection to detect and display the speed limit signs, Bump ahead, Railway Crossing, Oneway Traffic, No entery Vehicles, etc.

For this purpose we used transformer model and under that we are used Vit abbreviated form of vision transformers.

2. Transformer

Transformer is proved to be a simple and scalable framework for computer vision tasks like image recognition, classification, and segmentation, or just learning the global image representations. It demonstrated significant advantage in training efficiency when compared with traditional methods.

Transformer is Seq2Seq model, it has an encoder and a decoder. Seq2Seq models are particularly good at translation, where the sequence of words from one language is transformed into a sequence of different words in another language. A popular choice for this type of model is Long-Short-Term-Memory (LSTM)-based models. With sequence-dependent data, the LSTM modules can give meaning to the sequence while remembering (or forgetting) the parts it finds important (or unimportant). Sentences, for example, are sequence-dependent since the order of the words is crucial for understanding the sentence.Transformer is based on attention and self-attention.

The Encoder takes the input sequence and maps it into a higher dimensional space (n-dimensional vector). That abstract vector is fed into the Decoder which turns it into an output sequence. The output sequence can be in another language, symbols, a copy of the input, etc.

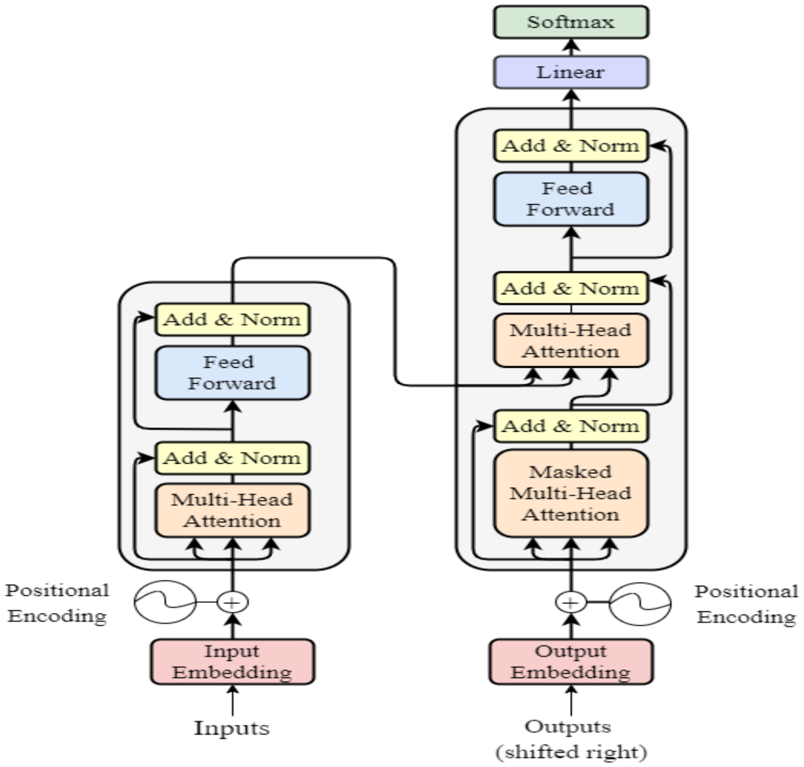
Both encoder and decoder are comprised of modules that can speak onto the top of each other multiple times. So what happens is the inputs and outputs are first embedded into n-dimension space, since we cannot use this directly. So we obviously have to encode our inputs, whatever we are providing. One slight, but important part of this model is positional and coding of different words. Since we have no recurrent neural network that can remember how to sequence is fed into the model, we need to somehow give every word or part of a sequence, a relative position since a sequence depends on the order of the elements. These positions are added to the embedded representation of each word. So this was a brief about Transformers.

2.1 Architecture Of Transformer

The architecture for image classification is the most common and uses only the Transformer Encoder in order to transform the various input tokens. However, there are also other applications in which the decoder part of the traditional Transformer Architecture is also used.

The Encoder is on the left and the Decoder is on the right. Both Encoder and Decoder are composed of modules that can be stacked on top of each other multiple times, which is described by *Nx* in the figure. We see that the modules consist mainly of Multi-Head Attention and Feed Forward layers. The inputs and outputs (target sentences) are first embedded into an n-dimensional space since we cannot use strings directly.

One slight but important part of the model is the positional encoding of the different words. Since we have no recurrent networks that can remember how sequences are fed into a model, we need to somehow give every word/part in our sequence a relative position since a sequence depends on the order of its elements. These positions are added to the embedded representation (n-dimensional vector) of each word.



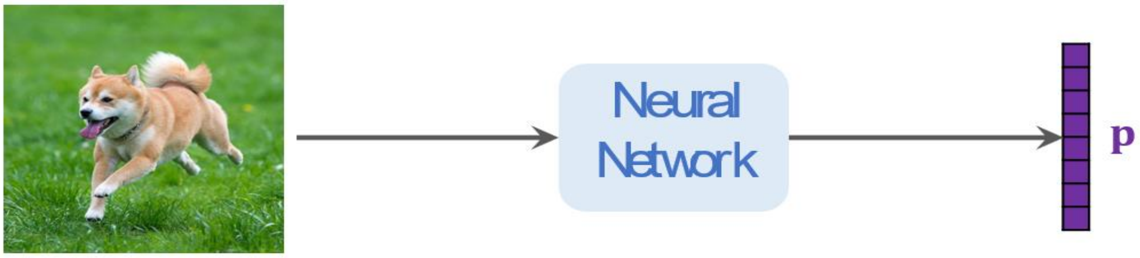
Transformer is very sufficient for large scale data. ViT demonstrates excellent performance when trained on sufficient data.Vision Transformer(ViT) can achieve excellent results with pure transformer architecture applied directly to a sequence of image patches for classification tasks.

3. Vision Transformers(ViT module)

Transformers are initially used in applications related to natural language processing (NLP) tasks. In computer vision research, there has recently been a rise in interest in Vision Transformers (ViTs) and Multilayer perceptrons (MLPs).

The Vision Transformer (ViT) model was introduced in a research paper published as a conference paper at ICLR 2021 titled “An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale”. It was developed and published by Neil Houlsby, Alexey Dosovitskiy, and 10 more authors of the Google Research Brain Team.

Transformer was developed in 2017 for natural languague processing. ViT is succesfull appication of transformer in computer vision but vit model does not have novelty .ViT is exactly the encoder network of transformer.



3.1 Working of Vision Transformer

The performance of a vision transformer model depends on decisions such as that of the optimizer, network depth, and dataset-specific hyperparameters. Compared to ViT, CNNs are easier to optimize.

The disparity on a pure transformer is to marry a transformer to a CNN front end. The usual ViT stem leverages a 16*16 convolution with a 16 stride. In comparison, a 3*3 convolution with stride 2 increases the stability and elevates precision.

CNN turns basic pixels into a feature map. Later, the feature map is translated by a tokenizer into a sequence of tokens that are then inputted into the transformer. The transformer then applies the attention technique to create a sequence of output tokens. Eventually, a projector reconnects the output tokens to the feature map. The latter allows the examination to navigate potentially crucial pixel-level details. This thereby lowers the number of tokens that need to be studied, lowering costs significantly.

Particularly, if the ViT model is trained on huge datasets that are over 14M images, it can outperform the CNNs. If not, the best option is to stick to ResNet or EfficientNet. The vision transformer model is trained on a huge dataset even before the process of fine-tuning. The only change is to disregard the MLP layer and add a new D times KD*K layer, where K is the number of classes of the small dataset.

3.2 Real world applications of vit transformers

Vision transformers have extensive applications in popular image recognition tasks such as object detection, segmentation, image classification, and action recognition. Moreover, ViTs are applied in generative modeling and multi-model tasks, including visual grounding, visual-question answering, and visual reasoning.

Video forecasting and activity recognition are all parts of video processing that require ViT. Moreover, image enhancement, colorization, and image super-resolution also use ViT models. Last but not least, ViTs has numerous applications in 3D analysis, such as segmentation and point cloud classification.

4. Dataset

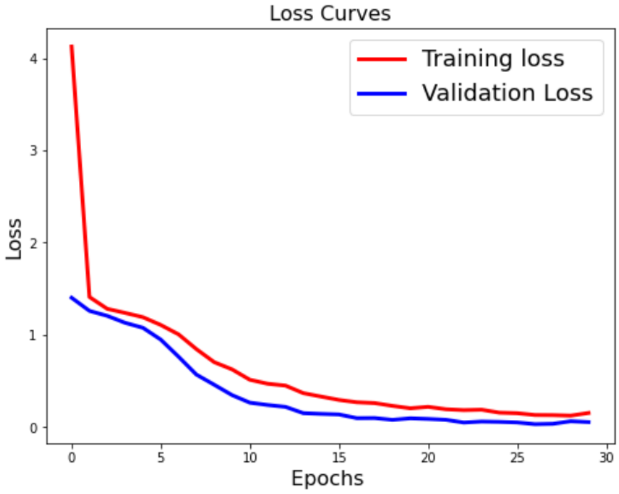
Class : 5

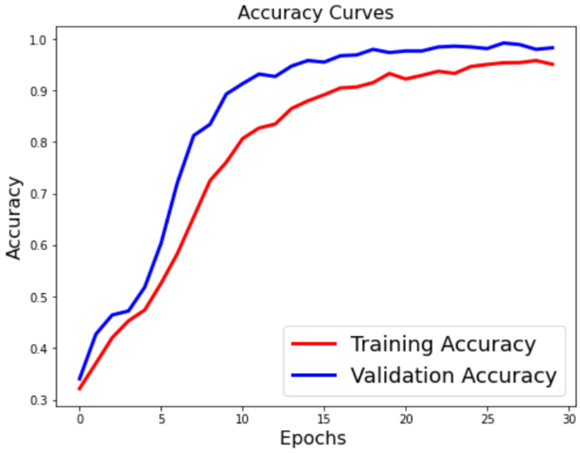
Class 0	Class 1	Class 2	Class 3	Class 4
X=79,Y=43 HEX #ffffff RGB (255,255,255) CMYK (0,0,0,0)	X=18,Y=66 HEX #696068 RGB (105,96,104) CMYK (0,9,1,59)	X=7,Y=38 HEX #2b2224 RGB (43,34,36) CMYK (0,21,16,83)	X=2,Y=22 HEX #191a1a RGB (25,26,26) CMYK (4,0,0,90)	X=34,Y=44 HEX #d19b52 RGB (209,155,82) CMYK (0,26,61,18)

4.1 Parameters of Dataset

```
learning_rate = 0.001
weight_decay = 0.0001
batch_size = 256
num_epochs = 30
image_size = 72
patch_size = 6
num_patches = (image_size // patch_size) ** 2
projection_dim = 64
num_heads = 4
transformer_layers = 8
mlp_head_units = [2048, 1024]
```

5. Result analysis





Test accuracy: 98.7%
Test top 5 accuracy: 100.0%

6. Conclusion

The traffic sign recognition is a very helpful driver assistance technique for increasing traffic and driver safety. In this project, low computing complexity, adaptive and accurate mechanisms have been applied to extract and recognize the content of traffic sign.The goal of this research is to and get good accuracy based on Traffic sign dataset. We shown promising result with respect to the accuracy of 98.7%.The recognition performance is evaluated by using Loss curve and Accuracy curve analysis.

The vision transformer model uses multi-head self-attention in Computer Vision without requiring the image-specific biases. The model splits the images into a series of positional embedding patches, which are processed by the transformer encoder. It does so to understand the local and global features that the image possesses.

Last but not least, the ViT has a higher precision rate on a large dataset with reduced training time.

7. References

https://thesai.org/Downloads/Volume7No1/Paper_93-Traffic_Sign_Detection_and_Recognition.pdf

https://www.hindawi.com/journals/mpe/2015/250461/

https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04

https://www.britannica.com/technology/transformer-electronics

https://theaisummer.com/transformer/

https://machinelearningmastery.com/the-transformer-model/

https://www.researchgate.net/publication/322839466_Deep_neural_network_for_traffic_sign_recognition_systems_An_analysis_of_spatial_transformers_and_stochastic_optimisation_methods

https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign

https://www.arcjournals.org/pdfs/ijrscse/v2-i6/6.pdf

https://www.researchgate.net/publication/344379511_RECOGNIZATION_OF_TRAFFIC_SIGN