Road traffic sign detection and classification using Capsule Network

Mr. Jean de Dieu TUGIRIMANA, Mr. Janvier RULINDA, Mr Antoine NZARAMBA, Mr Kotonko L. Tresor

ABSTRACT: Many years ago detecting and recognizing road traffic sign was done by the convolutional neural network (CNN). CNN has played an import role in the field of computer vision and a variant of CNN has proven to be successful in classification tasks across different domains; nevertheless, there are two disadvantages to CNN: one their failure to take into the consideration of important spatial hierarchies between features, and their lack of rotational invariance [1]. To overcome this concern, G. Hinton et al. propose a novel of neural network using the concept of capsules in a recent paper [1]. Our core idea is to use CapsNet to detect and do classification of road traffic sign. Capsule network has achieved the state-of-the-art accuracy of 98.3% on German traffic sign recognition Benchmark dataset.

Key terms: Capsule Network, Convolutional neural network, Road Traffic sign.

I. Introduction

Traffic scene analysis is a very important topic in computer vision and intelligent systems [2-7]. Road traffic signs are designed to inform drivers of the current condition and other important information on the road. Recently, deep learning methods have shown superior performance for many tasks such as image classification. One particular variant of deep neural networks, capsule networks have shown their strengths for task, including image classification, localization and detection.

Generally, CNN is used for all the state of the art deep learning neural network algorithms in most of the image related tasks. Convolution captures the spatial information of the image using the kernel function in convolution layer. A CNN consists of input, output and hidden layers. The hidden layer further consists of convolutional, pooling, fully connected and normalization layers. CNNs perform very well for image related operations, but they have some fundamental limits and disadvantages. The CNN fails [8] to capture the relative spatial and orientation relationships. CNN can get confused easily by image orientation or by change in pose.

Pose information can be rotated, thickness, skew, precise object position. Max polling is the biggest drawback for the CNN as they cannot propagate the spatial hierarchies between simple and complex objects which lead to invariance and makes them fail to capture the pose and the spatial relationships between the pixels in the image. CNN uses the max pooling layer which down samples the data and reduces the spatial information of data that is passed to the next layer. To address this drawback capsnet architecture is invented which reached the state of the art performance on the MNIST digits dataset and obtained better results than CNNs on Multi MNIST dataset.

II.RELATED WORK

Over the last decade, research in road traffic sign detection and classification has grown rapidly. A large number of novel ideas and effective methods have been proposed. Usually, the detection part hunts potential regions of traffic signs and the recognition part, determines the category of traffic signs

II.A. Traffic Sign Classification

Before the widespread adoption of capsule networks, various object detection methods were adapted for traffic-sign classification, e.g. based on SVMs [9] and sparse representations [10]. Recent capsule network approaches have been shown to outperform such simple classifiers when tested on the German Traffic Sign Recognition Benchmark dataset (GTSRB)

II.B. Using computer vision feature extraction methods

This is one of the earliest approaches under which a lot of algorithms and methodologies were proposed by the computer vision scientists before the advent of machine learning. One of the popularized techniques like HOG (Histogram Oriented Gradients)[11] was used for detection of pedestrians. In this method gradients of color image are computed along with different normalized, weighted histograms.

II.C. Using machine learning

Many kinds of machine learning algorithms like linear discriminant analysis(LDA)[13], random forest and kd-trees, ensemble classifiers, support vector machines (SVM) were used for the road traffic sign classification.

SVM is a classification algorithm which classifiers the data by dividing the n dimensional data plane with the hyper plane for classification [12]. SVM can even separate non-linear scattered data by transforming the classification plane to higher dimensions using non-linear kernel function which uses a method called kernel trick for its implementation.

LDA is based on maximum posteriori estimation of the class membership. Class densities are assumed to have multi variate Gaussian and common co-variance matrix.

Random Forest is an ensemble classifier method [14] which is based on the set of non-pruned random decision trees. Each decision tree is built using the randomly took training data. For classification testing data are evaluated by all the decision trees and the classification output is based on majority voting considering decisions of all the majority decision trees.

- *Mr. Jean de Dieu Tugirimana: currently pursuing master's degree program in Computer Science in Central South University, China, PH-(+86)15773172105. E-mail: titijado@gmail.com*
- *Mr. Janvier Rulinda: currently pursuing master's program in Computer Science in Central South University, China*
- *Mr Antoine Nzaramba :graduated masters student in computer science at Central South University, China*
- *Mr Kotonko Lumang amanga Tresor: Currently pursuing master's degree program in Computer Science in Central South University, China*

II.D. Using deep learning

To overcome the disadvantages of above mentioned conventional methodologies, new implementations based on deep learning algorithms replaced the previous methods [15] in recent years with increase in computing power and availability of standardized data sets and access to huge amounts of data. Convolutional neural networks are the state of the art algorithms achieving highest accuracy rate.LENET architecture [16] was the first CNN architecture for traffic sign classification.

Convolutional neural networks are biologically inspired multi-stage neural network architecture that learns the invariant features automatically. Each stage consists of filter bank (convolution) layer, non-linear transform layer, spatial pooling layer [17].The spatial pooling layer deceases the spatial information and acts like a complex cells in visual cortex. A gradient descent based optimizer is used for training and updating each filter to minimize loss function. The output of all the layers is fed to the classifier for improving the accuracy of classification.

III. DATA SET

German Traffic Sign Recognition Benchmark (GTSRB) dataset was described and visualized. It is a publicly available dataset and it is created from 10 hours video of driving on different roads in Germany. Video is captured using Prosilica GC 1380CH camera with frame rate of 25 fps and the traffic sign extraction is performed using the NISYS Advanced Development and Analysis Framework (ADAF)[18] module based software system.

After cleaning and removing the redundant and repetitive frames the dataset is reduced to total 51,840 images of the 43 classes. All the images in the dataset have 32*32 size and the total dataset is divided into training data and testing data. Total 39,209 images are present as training data and 12,630 images as test data

Fig 1. Sample images from GTSRB dataset

IV. Overview of Capsule Networks

Capsule networks represent a recent breakthrough in neural network architectures. They achieve state of the art accuracy on the MNIST dataset, a feat achieved traditionally by deep convolutional neural network architectures. Capsule networks introduce an alternative to translational invariance other than pooling through the use of modules, or capsules. Two key features distinguish them from CNN's: layer-based squashing and dynamic routing. Whereas CNN's have their individual neuron's squashed through nonlinearities, capsule networks have their output squashed as an entire vector. Capsules replace the scalar-output feature detectors

of CNNs with vector-output capsules and max-pooling with routing-by-agreement. Capsnet architectures typically include several convolution layers, with a capsule layer in the final layer.

IV.1 Dynamic Routing

Capsules output a vector, which means that it is possible to selectively choose which parent in the layer above the capsule is sent to. For each potential parent, the capsule network can increase or decrease the connection strength. This routing by agreement is much more effective at adding invariance than the primitive routing introduced by max-pooling.

IV.2. Reconstruction Regularization

Whereas traditional CNN's prevent over-fitting by using dropout, Capsule networks are regularized with a reconstruction auto-encoder. During training, all activity vectors are masked except for the activity vector corresponding to the correct digit. This activity vector is then used to reconstruct the input image. The output of the digit is then used to compute the loss. This encourages the network to learn a more general representation of images.

IV.3 Capsule Net Architecture

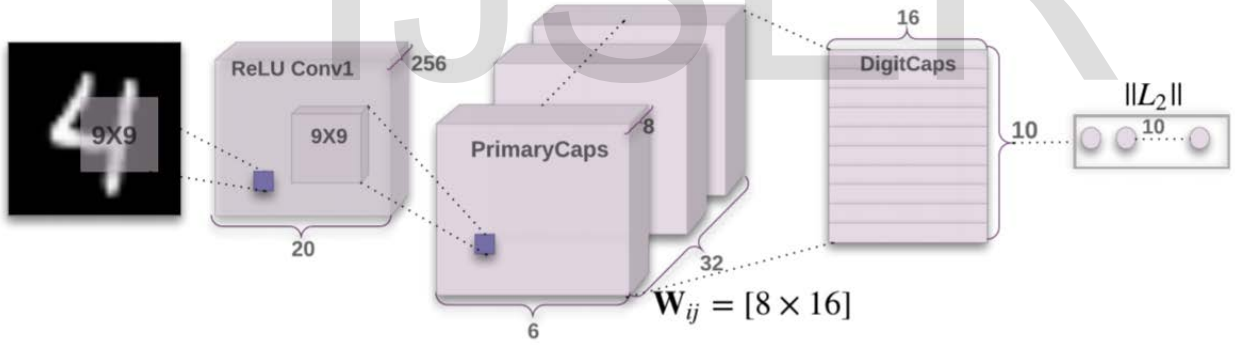

Fig 2.The Architecture of CapsNet from Original paper "Dynamic Routing between Capsules" by Sara Sabour, Nicholas Frosst and Geoffrey E. Hinton

Capsule networks consist of capsules rather than neurons. Capsule is a group artificial neural networks that perform complicated internal computations on their inputs and encapsulate the results in a small vector. Each capsule captures the relative position of the object and if the object pose is changed then the output vector orientation is changed [19] accordingly making them equi-variant.

Caps Net consists of multiple layers and the first layer is called primary capsules where each capsule receives a small part of the receptive field as input and tries to detect the pose of particular pattern. The output of the capsule is a vector and dynamic routing technique was used to ensure that the output is sent to the appropriate parent in the layer which can be inferred from fig. 3.

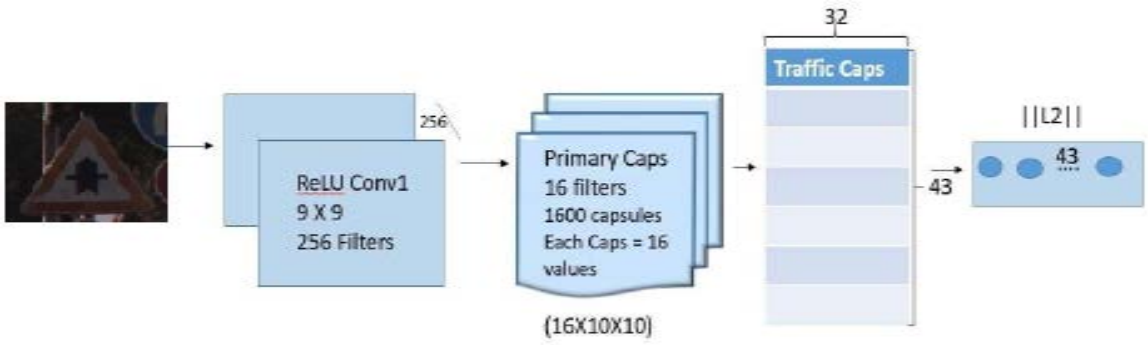

Fig 3. Capsule Sign architecture for road Traffic detection

IV.3.A. Computation of capsule vector inputs and outputs

The capsule computes a prediction vector [20] by multiplying the weight matrix(Wij) with its own output vector(ui). The coupling coefficient of that capsule corresponding output increases the scalar product and prediction [20] for that particular capsule output.

ûj|i = Wijui where **ûj|i** =prediction vector, **Wij**=weight matrix and **ui**=output vector.

IV.3.B. Squash Function

In capsule networks a non-linear activation function called squashing function [21] is used. This function converts the length of the output vector into the probability of the capsule connecting to that object. It performs shrinking of the long output vectors slightly below length one and short output vectors almost close to zero

$$vj = \frac{||Sj||^2}{1 + ||Sj||^2} \frac{Sj}{||Sj||}$$

Where sj=Total Input, vj=Vector Output of capsule j.

IV.3.C. Routing Algorithm

With vector output capsules[19] and max pooling with route by agreement. The dimensionality of the capsule increases with the increase In hierarchy because of the shift from place coding (encoded in continuous space) to rate coding (encoded in) and the high level capsules represent entities which are complex and have more degree of freedom. This route by agreement is efficient than the max pooling used in the CNNs.

$$S_j = \sum c_{ij} \hat{u}_{j|i}$$

Where **Sj**=summation matrix, **uj|i**=prediction vector, and **cij**=coupling coefficients determined by iterative dynamic routing.

V. EXPERIMENTS

V.A. Data collection

The data is come from hard disk in pickled format. Pickling is the process of saving a file in a serialized format before writing it into the disk. The total image is 25000 in training dataset and in testing dataset the total image is 11000.

The image brightness is enhanced with random uniform distribution of 0.6 to 1.5 and image contrast is also enhanced with random uniform distribution of 0.6 to 1.5. The size of image is 32*32. The training dataset is augmented by five-fold by replicating the available with data rotation±20,shear range of 0.2, width shift range of 0.2,horizontal flip which increased the training dataset size leading to better performance and regularizing thus avoiding the over fitting problem.

V.B. Network architecture

The architecture used for the road traffic sign detection consists of the input layer and initially convolutional layers as part of primary capsules and the output vector of primary capsule is sent to road traffic sign capsules.

- Input Layer: the input layer consists of input training images and the dimension is equal to the total training images.
- Primary Capsule Layer: the first layer that follows the input layer is the primary capsule layer and for calculating the output first two convolution layers were used. The first convolution layer consists of kernel size 9 and 256 filters and padding was not used. Rectified Linear unit was used as the non-linear activation function and a drop out of 0.7 which is fixed to be optimal after testing with different values.

  Output is reshaped to get the output vectors of primary capsules. Since the primary capsule layer is fully connected to the road traffic sign capsule layer the output vectors have to be squashed using the squashing function. Small epsilon value is added to the squash function to avoid the vanishing gradient problem while training. Now the output of the squash function is fed to the road traffic sign capsule layer.
- Road traffic sign capsule layer: to compute the output of road traffic sign capsules, calculate the predicted output vectors for each and every primary, traffic sign capsule pair and implement the route by agreement algorithm

  The road traffic sign capsule layer consist of 43 capsules each representing a particular class of the German traffic sign dataset with size of 32each. For each capsule i in the first layer predict the corresponding weights and output vectors of every capsule j in the second layer.

V.C. Reconstruction

A decoder network is added to the road traffic sign capsule network which consists of fully connected network layer which helps in the reconstruction of the input images by tuning the output of the road traffic sign capsule network. This feedback mechanism will make the network to preserve the information required for the reconstruction of the road traffic sign across the entire network. This acts as regularization and this avoids the over-fitting of the data and helps in proper generalization of road traffic signs.

- Decoder: the decoder consists of a non-linear activation layer of ReLU followed by a sigmoid activation layer.
- Mask: masking function is used to avoid the other entire output vector during training phase. And the reconstruction mask is realized using the one-hot function. For the target class its value will be one and for all the other classes its value will be zero.

V.D. The Losses

- Final loss: the final loss is the sum of margin loss and reconstruction loss scaled to a factor λ which acts as a scaling factor and it should be very much less than one

  F=(Margin Loss)-λ(Reconstruction Loss)

  Where F=Final Loss, λ=0.0005

  Margin loss should always dominate the Reconstruction loss in comparison. If reconstruction loss is more in the final loss then the model tries to exactly match output image with the input image of training dataset which lead to overfitting of the model to the training data
- Margin loss: the length of the instantiation output vector represent the probability of the respective capsule's entity exists or not. The digit class k has the longest vector output only if that road traffic sign is present in the input image.

  For every road traffic sign capsule k the margin loss is separate and it is given as

  $L_k=T_k\max(0,m^+-||vk||)^2 + \lambda(1-T_k)\max(0,||vk||-m^-)^2$

  The value of T_k is 1 if a traffic sign of class k is present and here m+=0.9 and m-=0.1.λ is a regularization parameter which stops the learning from shrinking the activity vector of all road traffic sign capsules.
- Reconstruction loss: it is the difference between the squares of the input image and reconstructed image

  $R=(\text{Input image})^2-(\text{Reconstructed image})^2$

  where R=Reconstruction loss

V.E. The Results

The model is evaluated using the testing data set of 11000 testing images. Accuracy is computed as the ratio of the correctly identified road traffic signs by the total number of road traffic signs [12]

$$Accuracy = \frac{\sum correctly\ identified\ road\ traffic\ signs}{Total\ number\ of\ road\ traffic\ signs}$$

With a batch size 30 obtained accuracy 98.3%

VI. CONCLUSION

Road traffic sign detection is a challenging task and capsule networks using their inherent ability to detect the pose and spatial variances perform better when compared to CNN's and capsule networks increase the reliability and accuracy by correctly performing image classification and recognition tasks even on indistinct, rotated and distorted images

VII. REFERENCES

[1] Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic Routing Between Capsules." arXiv preprint arXiv:1710.09829 (2017)

[2] Z. Zhang, T. Tan, K. Huang, Y. WangPractical camera calibration from moving objects for traffic scene surveillance IEEE Trans. Circuits Syst. Video Technol., 23 (3) (2013), pp. 518-533

[3] Z. Zhang, K. Huang, Y. Wang, M. LiView independent object classification by exploring scene consistency information for traffic scene surveillance Neurocomputing, 99 (2013), pp. 250-260

[4] C. Yao, X. Bai, W. Liu, L. Latecki, Human detection using learned part alphabet and pose dictionary, in: Proceedings of ECCV, 2014.

[5] C. Hu, X. Bai, L. Qi, X. Wang, G. Xue, L. MeiLearning discriminative pattern for real-time car brand recognition IEEE Trans. Intell. Transp. Syst., 16 (6) (2015), pp. 3170-3181

[6] Y. Xia, W. Xu, L. Zhang, X. Shi, K. MaoIntegrating 3d structure into traffic scene understanding with rgb-d data Neurocomputing, 151 (2015), pp. 700-709

[7] Q. Ling, J. Yan, F. Li, Y. ZhangA background modeling and foreground segmentation approach based on the feedback of moving objects in traffic surveillance systems Neurocomputing, 133 (2014), pp. 32-45

[8] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. ArXiv e-prints, June 2015

[9] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, and F. Lopez-Ferreras. Road-sign detection and recognition based on support vector machines. Intelligent Transportation Systems, IEEE Transactions on, 8(2):264–278, June 2007.

[10] K. Lu, Z. Ding, and S. Ge. Sparse-representation-based graph embedding for traffic sign recognition. IEEE Transactions on Intelligent Transportation Systems, 13(4):1515– 1524, 2012.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) /, volume 1, pages 886-893 vol. 1, June 2005

[12] Jung- Guk Park and Kyung-Joong Kim. Design of a visual perception model with edge-adaptive gabor filter and support vector machine for traffic sign detection. Expert Systems with Applications, 40(9):3679-3687,2013

[13] Yihui Wu, Yulong Liu, Jianmin Li, Huaping Liu, and Xiaolin Hu. Traffic sign detection based on convolutional neural networks. In neural networks (IJCNN), The 2013 International Joint Conference on pages 1-7. IEEE, 2013

[14] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In Neural Networks(IJCNN), The 2011 International joint conference on, paages 1453-1460. IEEE, 2011.

[15] C. K. Chandni, V. V. S. Variyar, and K. Guruvayurappan. Vision based closed loop pid controller design and implementation for autonomous car. In 2017 International Conference on Advances in computing communications and informatics(ICACCI), pages 1928-1933, Sept 2017

[16] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks, 32:323-332, 2012

[17] Pierre Sermanet and Yann LeCun. Traffic sign recognition with multiscale convolutional networks. In IJCNN, pages 2809-2813 IEEE, 2011

[18] Sebastian Houben, Johonnes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In international joint conference on neural networks,number 1288,2013

[19] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In international conference on Artificial Neural Networks, pages 44-51. Springer, 2011

[20] Dilin Wang and Qiang Liu. An optimization view on dynamic routing between capsule 2018

[21] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In Advances in Neural information processing systems, pages 3859-3869,2017