Investigation on the Effect of the Feature Extraction Backbone for Small Object Segmentation using Fully Convolutional Neural Network in Traffic Signs Application

H M Elhawary¹*, M I Shapiai², A Elfakharany²

¹ Dep. of Computer Engineering, Arab Academy for Science and Technology, Cairo, Egypt.  
² Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia.

*Corresponding author: hossam@aast.edu

Abstract. Small objects are widely found in different applications such as traffic signs and to segment those objects make it difficult to extract features due to the small number of pixels. Previous research has been done to show how error prone the semantic segmentation networks to small objects in variance of application such as medical images and remote sensing and how it leads to class imbalance. However, small object segmentation seems to be tricky and making the network struggle. Recently there are small amount of research has been done in the effect of the feature extraction backbone to the small object datasets. In this paper we investigate the effect of different backbone feature extraction such as AlexNet, VGGNet, GoogleNet on an imbalanced small objects dataset after grouping them by shape and colour in the Fully Convolutional Networks (FCN). We measure the performance on PASCAL VOC and Malaysian Traffic Sign Dataset (MTSD) showing the pixel accuracy, mean accuracy per class, mean IoU and frequency weighted IoU for each backbone and FCN. The results show that VGGNet as a backbone with Cross Entropy (CE) combined with Dice Loss (DL) achieves the highest score in mean IoU for imbalanced dataset but not for balanced dataset. However, in the imbalanced dataset major classes have a higher probability to confuse with minor classes due to the class imbalance. In conclusion we investigate different backbone networks with grouped labels dataset in shape and colour and we recommend using VGGNet FCN with CE combined with DL for imbalanced datasets.

1. Introduction

Over the past few years Convolution Neural Networks (CNN) become so popular for feature extraction. CNN-based models achieving state-of-the-art results in classification, localization and semantic segmentation. However, these models need big amount of data to achieve satisfied results. Not only the problem is in collecting the data but also there is the problem of imbalance between the classes for a multi-class dataset which can cause the major classes to have better accuracy than the minor classes. Small objects can cause difficulties in extracting the features due to the small amount of information [1]–[3].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
Most of the semantic segmentation networks such as FCN [4] use up-sampling deconvolution layers after the feature extraction part to predict the segmented object pixel-wise. The deconvolution layers are being affected by the feature extraction part and that’s why it is important to choose the right backbone network.

The rest of the paper is organized as the following: section 2 discusses some related work to solve the problem of imbalanced dataset and small objects, we introduce the framework and the metrics that we are going to use to investigate on Malaysian Traffic Sign Detection MTSD [5] dataset in section 3, and finally discuss our experimental results in section 4.

2. Deep Learning for Imbalanced and Small Objects

There are few deep learning techniques have been published to solve various real engineering applications. In traffic sign application, detection network [6]–[17], segmentation network [18]–[21], have been employed to detect traffic sign which known to be small in object and highly imbalance in nature. Most of the works mainly focus on the network architecture and augmentation [8], [11], [22], [23]. For instance, in biomedical field, the lesion is small and in grayscale requires modification in network architecture. Drozdzal et al [24] studied the effects of different kinds of skip connections on FCN with ResNet as a backbone for image segmentation in the biomedical field. They trained a very deep FCN that yielded good results on the EM dataset without the need for post-processing.

Besides, network architecture and augmentation technique, several authors have highlighted the importance of loss function. However, only few papers are discussed on the loss function implementation [11], [25] which found to offer better performance accuracy especially in dealing with small and imbalanced dataset. Ma et al [25] investigated the different loss functions used in CNN training for the segmentation task. They sorted the loss functions into a taxonomy of four categories: distribution-based losses, region-based losses, boundary-based losses and compound losses. They also concluded that for a mildly imbalanced dataset a Dice loss or generalized dice loss can yield a good performance, while a highly imbalanced dataset they recommend using a compound loss function such as: the exponential logarithmic loss function, the sum of Dice loss and cross entropy or combining boundary loss with generalized Dice loss. The author also highlighted on distribution-based losses known as WCE to solve the imbalance issue on the given dataset.

Since Ma et al [25] conclude that using Dice loss (DL) with cross entropy (CE) and conventionally WCE will offer better performance accuracy in dealing with imbalanced dataset. To the best of author knowledge, there is no investigation that has been carried out to examine the mentioned loss function for the proposed deep learning architecture for traffic sign detection.

3. Methodology

In this section we describe the setup that we have used to investigate the FCN with different backbones and loss functions moreover the metrics that we use to test each experiment accuracy on each dataset. The proposed methodology is mainly based on Shelhamer et al [4] architecture on FCN where they used pretrained convolutional layers of known classification networks (Alexnet, VGG or GoogLenet) and fine-tuned them, then they attached upsampling layers at the end to produce a segmentation map output with the same spatial size of the input.

3.1. Proposed framework

To make our investigation on FCN we have used the following framework which consist of five main parts as shown in Figure 1:
3.1.1. Input
The MTSD dataset [5] is a detection dataset which contains boundary box and class id for each object. Hence, we make use of the traffic sign shapes to create the segmentation ground-truth for the dataset as shown in Figure 2. In order to reduce the number of classes, we group the signs based on their shapes and colour as A. Madani proposed in [5] the frequency distribution after grouping are shown in Figure 3.
As shown in Figure 3 the dataset is highly imbalanced especially class Yellow Diamond and Red Circle are the major classes.

3.1.2. Backbone with deconvolution

The Input of the network is $C \times H \times W$ – where $C$ is 3-dimensional representing R, G and B. in general, the size either for $H$ or $W$ is capped to 800 pixels.

In this paper we have used two different backbone networks in our investigation, VGG16 which have five pooling layers as stated in Karen Simonyan et al. [26] and AlexNet as stated in Alex Krizhevsky et al. [27] We pretrained all the backbone with ImageNet [28]. The last two fully connected layers has been removed and replaced by 1x1 convolution with channel dimension number of classes to predict the score for each class including the background followed by deconvolution layer to bilinearly upsample of stride 32 to produce the pixel-wise output as stated in [4] the FCN32s.

3.1.3. Investigated Loss Function

The loss functions that being used are the following:

1. Cross-entropy (CE) as in Equation (1)
2. Dice Loss (DL) as in Equation (2)
3. Weighted Cross-entropy (WCE) as in Equation (3)

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} t_i^c \log(\text{softmax}(p_i^c))$$  \hspace{1cm} (1)

$$L_{Dice} = 1 - \frac{1}{C} \times \frac{2 \times \sum_{i=1}^{N} \sum_{c=1}^{C} t_i^c \ast \text{softmax}(p_i^c)}{\sum_{i=1}^{N} \sum_{c=1}^{C} t_i^c + \sum_{i=1}^{N} \sum_{c=1}^{C} p_i^c}$$ \hspace{1cm} (2)

$$L_{WCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c t_i^c \log(\text{softmax}(p_i^c))$$ \hspace{1cm} (3)

In this investigation, $w_c$ is defined based on the following equation is employed

$$w_c = 1 - \frac{f_c}{\sum_{i=1}^{C} f_i}$$ \hspace{1cm} (4)

Where $t_i^c$ is target label of class $c$ corresponding to pixel $i$. $p_i^c$ is the model prediction of class $c$ of pixel $i$. $w_c$ is weight of class $c$, $C$ is the total number of classes and $N$ is the total number of pixels.
3.2. Performance Measure

In order to validate the performance of the investigated loss function in the FCN network, the following performance index are used as used in [4]

1. **Mean IoU**: the mean Intersection over Union (IoU).

2. **Accuracy**: accuracy per pixel.

3. **Accuracy per class**: mean accuracy per class.

4. **fwavacc**: frequency weighted Intersection over Union (IoU).

4. Results and Discussion

4.1. Results

As shown in Table 1 the validation results for the MTSD shows the *four metrics* that has been discussed in section 3.2. In general, there are four different investigation Loss function that embedded in the FCN network. The best performance based on Mean IoU is CE + DL followed WCE which offers at 59.99%, 59.71% respectively. The second metric that provides insight of the prediction is accuracy per class. In this investigation, WCE outperforms the other loss function. However, accuracy and fwavacc for all investigated loss function are always 99%

| Model          | Loss Function | Learning rate | epoch | iteration | Mean IoU | Accuracy | Accuracy per class | fwavacc  |
|----------------|---------------|---------------|-------|-----------|----------|----------|--------------------|----------|
| VGG16-FCN32s   | 1e-5*CE + DL  | 1.0e-05       | 97    | 88000     | 59.99%   | 99.86%   | 74.24%             | 99.759715% |
| VGG16-FCN32s   | CE            | 1.0e-10       | 93    | 84000     | 54.64%   | 99.86%   | 67.80%             | 99.74%   |
| VGG16-FCN32s   | WCE           | 1.0e-10       | 110   | 100000    | 59.71%   | 99.84%   | 76.43%             | 99.71%   |
| ALEXNET-FCN32s | CE            | 1.0e-10       | 110   | 100000    | 30.61%   | 99.78%   | 39.26%             | 99.62%   |

4.2. Discussion

We focus on Mean IoU and Accuracy per class for MTSD since the Accuracy and fwavacc are including the background pixels which in case of traffic signs datasets most of the images are background, objects are too small and that makes the accuracy irrelevant and very high always greater than 99%. Table 1 shows that CE loss function with learning rate of $10^{-10}$ reach best mean IoU in iteration 84000 with value
of 54.64% and 67.80% accuracy per class compared with AlexNet 30.61% mean IoU and 39.26% accuracy per class which indicate that VGG16 is giving better result than AlexNet.

On the other hand, to solve the imbalance problem we have tried the DL for the multi-class but it shows no learning at all but when we try the combined CE with DL loss functions it shows an increase in both mean IoU and Accuracy per class by 5.35% and 6.44% compared to use of CE loss only.

We use the formula in Equation (4) to give the classes with high frequency a lower weight and the classes with low frequency get higher weights. However, the WCE shows better Mean IoU than CE only and better Accuracy per class than CE and CE combined with DL which means that using DL with CE help to solve the class imbalance but it might require to choose different hyperparameters to enhance the accuracy per class results. A visualized sample results of combined loss function with the combination of DL with CE is shown in Figure 4.

![Figure 4. Sample result of iteration 88000 for combined loss function CE + DL](image)

| Model          | Loss Function | Learning rate | epoch | iteration | Mean IoU | Accuracy | Accuracy per class | fwavacc |
|----------------|---------------|---------------|-------|-----------|----------|----------|-------------------|---------|
| VGG16-FCN32s   | 1e-5*CE + DL  | 1.0e-10       | 49    | 72000     | 48.85%   | 86.09%   | 61.48%            | 76.90%  |
As shown in Table 2 the VOC dataset show that using the CE only is better than combining with the DL since the VOC is already balanced and does not have small objects. Hence, the investigated Dice with CE and WCE are found to work well only for imbalanced dataset.

5. Conclusion
In conclusion we compared the results of FCN 32 stride deconvolution with VGG16 and AlexNet backbone and different loss functions such as Cross entropy and Dice Loss to compare between the MTSD which have small objects and is highly imbalanced and the PASCAL VOC 2011 which doesn’t have the imbalance problem nor small objects. The results shows that Cross entropy can reach 62.36% mean IoU and accuracy per class of 74.04% with PASCAL VOC but the results are not satisfying when train using MTSD which have imbalance and small objects problem with 54.64% mean IoU and 67.80% accuracy per class. However, our investigation shows that combining the Dice Loss with Cross entropy loss increase the results by 5.35% in mean IoU and 6.44% in accuracy per class. We recommend using combination of Dice and Cross entropy when dealing with imbalanced data only.

References
[1] M. Kampffmeyer, A. B. Salberg, and R. Jenssen, 2016, “Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 680–688, doi: 10.1109/CVPRW.2016.90.
[2] A. Takeki, T. T. Trinh, R. Yoshihashi, R. Kawakami, M. Iida, and T. Naemura, 2016, “Detection of small birds in large images by combining a deep detector with semantic segmentation,” in Proceedings - International Conference on Image Processing, ICIP, 3977–3981, vol. 2016-Aug, doi: 10.1109/ICIP.2016.7533106.
[3] X. Y. Zhou, C. Riga, S. L. Lee, and G. Z. Yang, 2018, “Towards Automatic 3D Shape Instantiation for Deployed Stent Grafts: 2D Multiple-class and Class-imbalance Marker Segmentation with Equally-weighted Focal U-Net,” in IEEE International Conference on Intelligent Robots and Systems, 1261–1267, doi: 10.1109/IROS.2018.8594178.
[4] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation.”
[5] A. Madani and R. Yusof, 2016, “Malaysian Traffic Sign Dataset for Traffic Sign Detection and Recognition Systems,” J. Telecommun. Electron., vol. 8, no. 11.
[6] F. Shao, X. Wang, F. Meng, J. Zhu, D. Wang, and J. Dai, May 2019, “Improved Faster R-CNN Traffic Sign Detection Based on a Second Region of Interest and Highly Possible Regions Proposal Network,” Sensors (Basel)., vol. 19, no. 10, doi: 10.3390/s19102288.
[7] Z. Zhang, X. Zhou, S. Chan, S. Chen, and H. Liu, 2017, “Faster R-CNN for small traffic sign detection.””, in Communications in Computer and Information Science, 155–165, vol. 773, doi: 10.1007/978-981-10-7305-2_14.
[8] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, Jun. 2016, “Traffic-Sign Detection and Classification in the Wild,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2110–2118, doi: 10.1109/CVPR.2016.232.
[9] Y. Jin, Y. Fu, W. Wang, J. Guo, C. Ren, and X. Xiang, 2020, “Multi-Feature Fusion and
Enhancement Single Shot Detector for Traffic Sign Recognition,” *IEEE Access*, 38931–38940, vol. 8, doi: 10.1109/ACCESS.2020.2975828.

[10] Y. Lu, J. Lu, S. Zhang, and P. Hall, 2018, “Traffic signal detection and classification in street views using an attention model,” *Comput. Vis. Media*, 253–266, vol. 4, no. 3, doi: 10.1007/s41095-018-0116-x.

[11] U. Kamal, T. I. Tommoy, S. Das, and M. K. Hasan, Apr. 2020, “Automatic Traffic Sign Detection and Recognition Using SegU-Net and a Modified Tversky Loss Function with L1-Constraint,” *IEEE Trans. Intell. Transp. Syst.*, 1467–1479, vol. 21, no. 4, doi: 10.1109/TITS.2019.2911727.

[12] P. Cheng, W. Liu, Y. Zhang, and H. Ma, 2018, “LOCO: Local Context Based Faster R-CNN for Small Traffic Sign Detection,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 329–341, vol. 10704 LNCS, doi: 10.1007/978-3-319-73603-7_27.

[13] T. Yang, X. Long, A. K. Sangaiah, Z. Zheng, and C. Tong, May 2018, “Deep detection network for real-life traffic sign in vehicular networks,” *Comput. Networks*, 95–104, vol. 136, doi: 10.1016/J.COMNET.2018.02.026.

[14] H. S. Lee and K. Kim, May 2018, “Simultaneous Traffic Sign Detection and Boundary Estimation Using Convolutional Neural Network,” *IEEE Trans. Intell. Transp. Syst.*, 1652–1663, vol. 19, no. 5, doi: 10.1109/TITS.2018.2801560.

[15] C. Han, G. Gao, and Y. Zhang, May 2019, “Real-time small traffic sign detection with revised faster-RCNN,” *Multimed. Tools Appl.*, 13263–13278, vol. 78, no. 10, doi: 10.1007/s11042-018-6428-0.

[16] Z. Ou, F. Xiao, B. Xiong, S. Shi, and M. Song, 2019, “FAMN: Feature Aggregation Multipath Network for Small Traffic Sign Detection,” *IEEE Access*, 178798–178810, vol. 7, doi: 10.1109/ACCESS.2019.2959015.

[17] Z. Liu, J. Du, F. Tian, and J. Wen, 2019, “MR-CNN: A Multi-Scale Region-Based Convolutional Neural Network for Small Traffic Sign Recognition,” *IEEE Access*, 57120–57128, vol. 7, doi: 10.1109/ACCESS.2019.2913882.

[18] H. Li, F. Sun, L. Liu, and L. Wang, Dec. 2015, “A novel traffic sign detection method via color segmentation and robust shape matching,” *Neurocomputing*, 77–88, vol. 169, doi: 10.1016/j.neucom.2014.12.111.

[19] S. Pei, F. Tang, Y. Ji, J. Fan, and Z. Ning, Apr. 2018, “Localized Traffic Sign Detection with Multi-scale Deconvolution Networks.”

[20] C. Timbus, V. Miclea, and C. Lemnaru, 2018, “Semantic segmentation-based traffic sign detection and recognition using deep learning techniques,” in 2018 IEEE 14TH INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTER COMMUNICATION AND PROCESSING (ICCP), 325–331.

[21] Y. Zhu, M. Liao, M. Yang, and W. Liu, Jan. 2018, “Cascaded Segmentation-Detection Networks for Text-Based Traffic Sign Detection,” *IEEE Trans. Intell. Transp. Syst.*, 209–219, vol. 19, no. 1, doi: 10.1109/TITS.2017.2768827.

[22] V. D. Hoang, M. H. Le, T. T. Tran, and V. H. Pham, 2018, “Improving Traffic Signs Recognition Based Region Proposal and Deep Neural Networks,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 604–613, vol. 10752 LNAI, doi: 10.1007/978-3-319-75420-8_57.

[23] L. Abdi and A. Meddeb, 2017, “Deep learning traffic sign detection, recognition and augmentation,” in Proceedings of the Symposium on Applied Computing - SAC ’17, 131–136, doi: 10.1145/3019612.3019643.

[24] M. Drozdal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, Aug. 2016, “The Importance of Skip Connections in Biomedical Image Segmentation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 179–187, vol. 10008 LNCS.

[25] J. Ma, May 2020, “Segmentation Loss Odyssey.”

[26] K. Simonyan and A. Zisserman, Sep. 2014, “Very Deep Convolutional Networks for Large-Scale
Image Recognition.”

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks.” 1097–1105, 2012.

[28] O. Russakovsky et al., Sep. 2015, “ImageNet Large Scale Visual Recognition Challenge,” Int. J. Comput. Vis., 211–252, vol. 115, no. 3, doi: 10.1007/s11263-015-0816-y.