Artificial Intelligence and Human-Computer Interaction
Y. Ye and P. Siarry (Eds.)

doi:10.3233/FAIA240147

A Semantic Segmentation Algorithm in Workshop Scenarios

Hongmei YIN [a], Gege JIN [b], Zhijun LI [c], Chen CHEN [d] and Jiyong HUA [c,1]
[a] *Yangzhou Customs, Yangzhou, China*
[b] *Jiangsu United Vocational and Technical College Yangzhou Technician Branch, Yangzhou, China*
[c] *Jiangsu College of Tourism, Yangzhou, China*
[d] *Yangzhou University, Yangzhou, China*

Abstract. The classification of workshop objectives based on deep learning is the foundation of intelligent workshop management. There are various types of targets in the workshop, with variable geometric shapes and disorderly distribution of targets. In this article, we propose a deep learning based solution that can improve the speed and accuracy of target classification. A deep neural network model can train specialized models for specific work environments. The experiment has shown that the scheme has significant improvements in accuracy and visual effects.

Keywords. Deep learning, Workshop, Classification

1. Introduction

In specific scenarios, the goals are diverse. The target classification algorithm is gradually maturing and recognizing scenes, landmarks, parts, etc. The context takes into account the spatial and temporal connections of different targets in the scene. Spatial context has proven to be very useful, and contextual connections can be established using perspective clues.

In addition to pixel level category labels, semantic regions should also be labeled with their shape or structure, hierarchically. An advanced boundary detection method can be developed to convert this data and other pixel level segmentation data into more useful forms.

Training and evaluation need t a large number of data samples. We create a workshop scene database, which comes from actual photos taken in the workshop. We label every image according to 11 target categories.

In order to quantitative evaluating various recognition, detection, and segmentation, we conducted test statistics on all images in the database. The statistical results indicate that our algorithm has achieved good results.

[1] Corresponding Author: Jiyong Hua, Yangzhou university, China. E-mail:576461158@qq.com.

2. Related works

Chen, Liang Chieh et al. proposed a semantic segmentation model for codecs with depth wise separable convolution [1]. Brostow, G. J. et al. [2] conducted detailed experiments on the CamVid dataset. Divide the targets into 32 categories and evaluate them from three perspectives: target classification, pedestrian detection, and label propagation.

In 2015, Long et al of the University of California, proposed the full convolutional neural network (FCN) based on the classic network [3]. This model abandons the fully connected layer and adds layers with spatial translation invariant forms such as up-sampling layer and deconvolution layer. Different from traditional segmentation methods based on image blocks, FCN proves that convolutional neural network under end-to-end and pixel-to-pixel training can significantly improve the computational efficiency and prediction performance of semantic segmentation. End-to-end training lays the foundation for the development of subsequent semantic segmentation algorithms.

The construction method of FCN is to adapt all fully connected layers of traditional convolutional networks into dense convolutional layers of corresponding sizes. Based on VGGNet, FCN has adapted the last three layers of the VGGNet network into 1 × A multi-channel convolutional layer with the same vector length. The entire network model is entirely composed of convolutional layers without vectors generated by fully connected layers.

Full convolutional network (FCN) uses convolutional neural network to realize the transformation from image pixels to pixel categories [4]. It is different from classic convolutional neural network. A fully convolutional network transforms the height and width of the middle feature layer back to the size of the input image. Transposed convolution layers make the predicted results correspond one-to-one with the input image in spatial dimensions.

Based on the fully convolutional network model [5], there are significant improvements on several segmentation benchmarks [6]. Some variant models utilize contextual information for segmentation [7][8]. The multi-scale input [9] orthogonal set adopt a probability graph model [10].

The feature maps of large and small prediction bounding boxes are different. To ensure that the input feature dimensions of FCN are the same, pyramid pooling used.

To solve the problem of inputting images of varying sizes into fully connected layers with constant length, it is necessary to convert feature maps of any size into feature vectors of fixed size. Models such as PSPNet [11] or DeepLab [12] perform pyramid pooling on multiple grid scales[13].

Depth wise separable convolution or group convolution is a powerful operation that can maintain similar performance while reducing computational costs and parameter size.

3. Proposed method

We control the resolution network of deep convolutional neural computation features and adjust the filter's field of view to capture multi-scale information. The standard convolution operation has extended. In the case of a two-dimensional signal, for each position i, and convolutional filter w, on the output feature map y, apply deep convolution on the input feature map x.

y[i]=

$$\sum_{k} x[i + r \cdot k]w[k] \tag{1}$$

The encoder features in DeepLabv3 usually calculated using an output step of 16. The feature was up-sampled 16 times twice. This is a decoder module. However, this decoding module may not be able to recover the details of object segmentation successfully. Therefore, we propose a simple and effective decoder module. The encoder features are first bilinear up-sampled with a factor of 4, and then cascaded with corresponding low-level features from the network backbone with the same spatial resolution. Appling another factor to low-level features 1× 1, convolution is used to reduce the number of channels. The corresponding low-level features typically contain a large number of channels. This may outweigh the importance of enriching encoder features and make training more difficult.

4. Experiments and evaluation

We took 200 images of various scenes in the workshop and used ImageLabeler to mark them manually. **Figure 1** shows the classification of source images. We classify common objects in the workshop into 11 categories: profiles, glass, assembly, window, fag, rack, shelf, product, bracket, screw, and metal. Each category corresponds to one color.

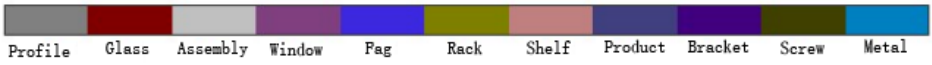

Figure 1. Class and corresponding color.

In our sample library, we have counted the number of pixels for various classifications. Showing as **Figure 2**.

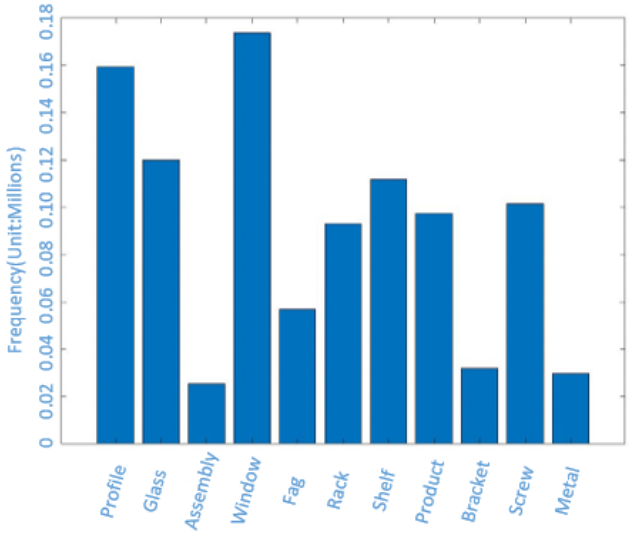

Figure 2. Neural network model.

Ideally, the number of pixels corresponding to each type should be roughly equal. In the pictures we collected, the number of assembly, fag, and bracket is relatively small. This imbalance can lead to uneven training. In training, we use different weights to adjust for this difference.

Our network structure diagram is as below in **Figure 3**.

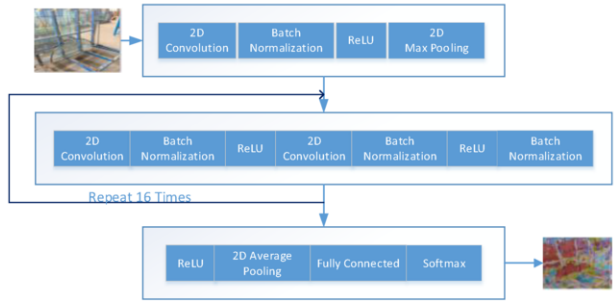

Figure 3. Neural network model.

The Learning rate is gradually declining, falling 0.2 every 10 epochs. Considering that we are training on a single CPU and have limited memory, we take the batch image size as 4. In order to utilize the sample data, we expanded the image. On the left and right sides of the image, add 10 pixels each.

Figure 4. Training procedure.

On AMD Ryzen 7 PRO 5845, 16GB memory, we trained for 45 minutes. Showing as **Figure 4**, the validation accuracy remained stable at 100%.**Figure 5** shows the visual effect.

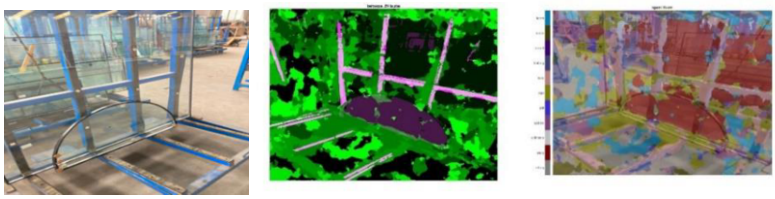

Figure 5. Original image, expected image, and result image.

The semantic segmentation results distinguish different targets well. The effectiveness of segmentation can have evaluated using the intersection over union matrix. Table 1 shows the statistic results.

Table 1 IoU evaluation.

Class	IoU
Profile	0.93628
Glass	0.87276
Assembly	0.57413
Window	0.96906
Fag	0.87979
Rack	0.94296
Shelf	0.7339
Product	0.76689
Bracket	0.76526
Screw	0.51442
Metal	0.79739

Statistical analysis conducted on all samples in the dataset. As **Table 2** shows, statistical results are based on the accuracy, IoU, and score corresponding to each class.

Table 2. Accuracy, IoU and MeanBFScore.

Class	Accuracy	IoU	MeanBFScore
Profile	0.94655	0.91702	0.91948
Glass	0.86472	0.84423	0.71884
Assembly	0.77632	0.33911	0.70544
Window	0.95286	0.94104	0.84832
Fag	0.90606	0.77672	0.79707
Rack	0.89632	0.79827	0.76612
Shelf	0.81612	0.51053	0.64256
Product	0.88429	0.5336	0.58259
Bracket	0.94219	0.83678	0.82065
Screw	0.8931	0.54522	0.72305
Metal	0.95257	0.67691	0.65654

Overall, the algorithm performs the best in detecting profiles. Other classifications, such as shelves and finished products, are not very effective. In the future study, the number of samples needs to be increased, and the trained model may have better classification performance.

5. Conclusion

The experimental results show that deep learning algorithms can achieve target segmentation in multi classification scenarios. To improve the accuracy of segmentation, manual annotation needs to be further precise. Increasing the number of training samples can also improve the equilibrium of the model. Enable the training results to balance each classification. Further work will have conducted in these two directions. We will streamline the model and remove some unnecessary hidden layers. We hope that this operation can shorten the training time while maintaining accuracy.

References

[1] Liang-Chieh Chen, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. ECCV (2018) 324-331.
[2] Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters. Vol. 30, Issue 2, 2009, p.88-97.
[3] Kohli, P., Torr, P.H., et al. Robust higher order potentials for enforcing label consistency. IJCV 82(3) (2009) 302–310.
[4] Pinheiro, P., Collobert, R. Recurrent convolutional neural networks for scene labeling. ICML (2014)655-662.
[5] Vemulapalli, R., Tuzel, O., Liu, M.Y., Chellappa, R.Gaussian conditional random field network for semantic segmentation. CVPR. (2016)413-419.
[6] Jian Yao, Sanja Fidler, Raquel Urtasun. Describing The Scene as A Whole: Joint Object Detection, Scene Classification and Semantic Segmentation[J], Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012: 702-709.
[7] Saining X, Ross B G, Piotr D, Zhuowen T, Kaiming H, et al. Aggregated Residual Transformations for Deep Neural Networks. [C], Computer Vision and Pattern Recognition, 2017, abs/1611.05431(): 5987-5995.
[8] Liang-Chieh C, George P, Iasonas K, Kevin M, Alan L Y, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. [J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.
[9] Stephen Gould, Richard Fulton, Daphne Koller. Decomposing a scene into geometric and semantically consistent regions[C], IEEE International Conference on Computer Vision, 2009(1): 1-8.
[10] Jampani, V., Kiefel, M., Gehler, P.V. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks. CVPR (2016)4452-4461.
[11] Farabet, C., Couprie, C., Najman, L., LeCun, Y. Learning hierarchical features for scene labeling. PAMI (2013)1915-1929.
[12] Abadi, M., Agarwal, A., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 (2016)
[13] LeCun, Y., Bottou, L., Bengio,Y., Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE (1998)2278-2324.