Pedestrian segmentation based on U-net.

Tom Hanks (Student ID: xxxxxxxxx)

Abstract

Image segmentation is a basic task of computer vision for image recognition and understanding. It's the pre-program for recognizing and understanding an image. The effect of image segmentation is self-evident for understanding the image. The accuracy and effectiveness of image segmentation are of great significance to the development of intelligent computer vision. This article will review some commonly used and popular image segmentation algorithms and will show pedestrian segmentation experiments project based on U-Net. Project source code available at my github channel[1].

1 Literature review

Earlier image segmentation algorithms were general based on cognition and basic knowledge on the pixel level, which is the similarity of pixels within objects and the dissimilarity of object boundaries. These common algorithms include thresholding, edge detection-based methods, Region Growing, Split and merge methods. The classic thresholding algorithm for two-segment, Otsu's algorithm is to exhaust the appropriate threshold based on statistical indicators, so that the internal pixel distance of the two parts of the segmentation is the smallest, and the distance between the two parts is the largest [1]. The edge detection-based algorithm always relies on the clear boundary between the objects in the image and the boundary line that can be connected. These traditional methods can perform well only on very simple tasks or special condition images and are difficult to competent for general image segmentation tasks.

Many machine learning (ML) algorithms are more versatile and accurate than previous traditional algorithms, often used for image segmentation. Image segmentation can be regarded as a classification problem, and the goal is to classify different pixels in one image into classes belonging to different objects. It should be noted that the classification model requires some labelled samples to learn the appropriate parameters for effective object segmentation. Numerous classification models such as Decision Tree (DT), Logistic regression, Support Vector Machines (SVM), Multi-layer Perceptron (MLP), etc. are all applied to various classification tasks [2]. Liu et al. [3] proposed a decision tree based semantic segmentation image retrieval system, which optimizes the tree pruning process and improves performance. Mizushima and Lu [4] demonstrated an effective algorithm for classifying and grading

[1] `https://github.com/StevenHuang2020/Pedestrian-Segmentation`

apples using image segmentation based on SVM and Otsu's thresholding method, with an error rate less than 2 percent. On the other hand, the clustering model can be selected to do image segmentation tasks when there are no labeled samples. Common clustering algorithms include K-Means, Agglomerative clustering, DBSCAN, OPTICS, etc. The fastest, simplest and efficient K-Means algorithm, due to its linear time complexity, is particularly suitable for clustering large-scale data sets. However, the clustering numbers parameters required by this algorithm must be determined through accurate measurements. Ray and Turi [5] proposed an algorithm based on distance measurement within and between clusters, which can automatically calculate the number of clusters.

Due to its high accuracy and strong versatility, Deep learning is the current new generation algorithms for generic image segmentation. Unlike the previous algorithms are basically based on the pixel level, the deep learning algorithm performs image segmentation rely on automatically extracted semantic features, which significantly increases robustness and accuracy. According to the research of Minaee et al. [6], the current classic CNN models for semantic segmentation include fully convolutional network (FCN), SegNet, U-Net, V-net, etc. The end-to-end fully convolutional network FCN obtains good segmentation results on the PASCAL VOC data set by changing the fully connected layer in the current excellent image classification neural network [7]. Noh et al. [8] proposed a convolutional network based on Encoder-Decoder that includes transpose convolution and unpooling layers, which alleviates the limitations of FCN. The other encoder-decoder model, the Segnet, which directly uses the pooling indices of the previous max-pooling layer in the encode upsampling stage, thereby effectively reducing the inference time[9]. Ronneberger, et al. [10] proposed the U-Net model inspired by the previous FCN and encoder-decoder architecture model and achieved good performance on the medical image segmentation task. Milletari et al. [11] proposed V-net, a fully CNN model based on U-net and successfully segment 3D MRI prostate medical images.

2 U-net Model

The experiment model is borrowed from the U-Net model and the architecture of the U-Net model is shown in Figure 1. The input of the model is modified to 255x255x1, and the output is a segmentation mask image of the same size. The model uses only convolutional layers and removes fully connected layers and uses convolution and deconvolution layers at different

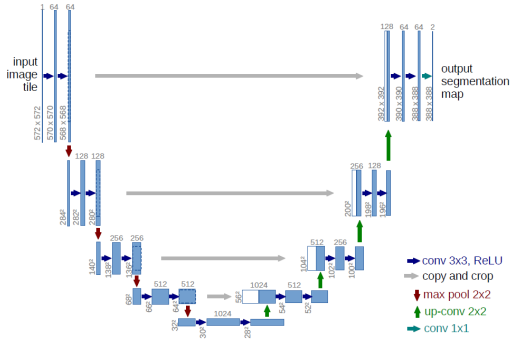


Figure 1: U-Net Architecture.

scales to increase the ability to recognize objects of different scales. The last convolutional layer uses sigmoid as the activation function. The loss function uses Dice loss which defined as follows,

$$Loss(y_{true}, y_{pred}) = 1 - \frac{2\sum_{pixels} y_{true} y_{pred} + smooth}{\sum_{pixels} y_{true}^2 + \sum_{pixels} y_{pred}^2 + smooth} \tag{1}$$

Where smooth=1, only used to avoid the denominator being equal to zero. The numerator part in the loss equation stands for the intersection of the prediction and the ground truth label, whereas the denominator part represents the union of these two vectors.

3 Dataset and Augmentation

The pedestrian benchmark Penn-Fudan dataset[2] will be used in my project. The data set contains 170 images, and each photo contains one or more pedestrian annotations. These photos were taken randomly from the road around the university campus. All images are three-channel colour images in PNG format, and the annotation information contains the bounding box and pedestrians' segmentation mask. Only segmentation mask information was used in this project.

It is difficult to drive the CNN model due to the sample sizes are relatively small, some appropriate methods must be used for image augmenta-

[2] `https://www.cis.upenn.edu/~jshi/ped_html/`

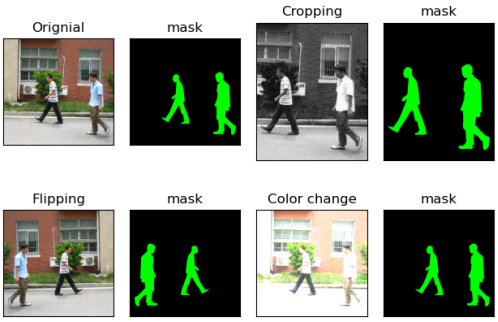
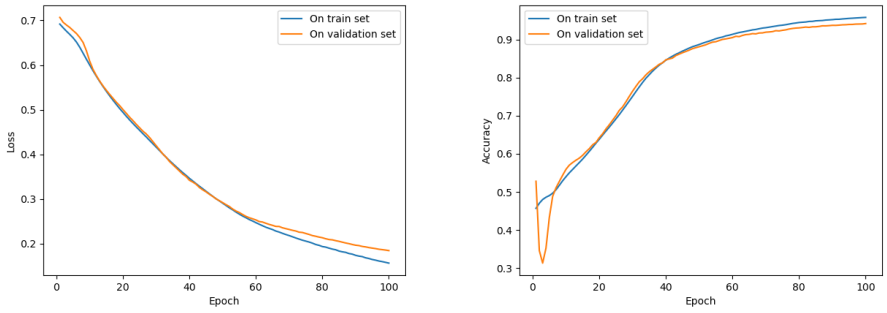


Figure 2: Augmentation by cropping, flipping, and colour conversion.

tion. These methods include resizing, cropping, flipping and colour conversion. The label segmentation mask needs to be modified at the same time while generating new images. Colour conversion using RGB channel separation, blur operation, brightness and contrast adjustment, gamma correction method, etc. Figure 2 shows different methods for image augmentation and the corresponding results. The number of images is expanded to 34680 from only 170 after using the above methods.

4 Training and Result analyses



(a) Loss. (b) Accuracy.

Figure 3: Loss and accuracy of the first 100 epochs.

The experiment was developed in python and OpenCV-python library and the model trained on Google Collaboratory (GPU). The augmented dataset is divided into two parts before training, 70 percent of them is the training

set, and the rest is the test set. Accuracy measurement uses intersection over union (IOU). Figure 3 shows the loss and accuracy of the first 100 epochs on the training set and the test set respectively. After thousands of epochs, loss on the training set are reached -0.2689 and the accuracy reached to 0.8978. Figure 4 and Figure 5 respectively demonstrates the pedestrian segmentation effect on different images.

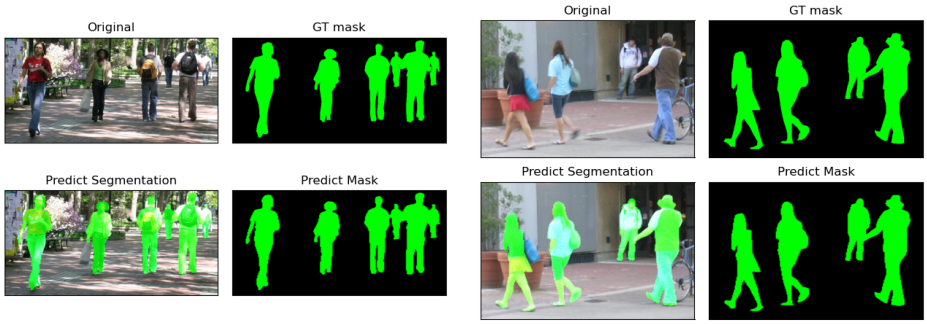


Figure 4: Detection effect of images form dataset.

Figure 5: Detection effect of images not in dataset.

5 Conclusion

According to the results of the experiment, 1) Segmentation for small pedestrians and overlapping pedestrians is not good. 2) The segmentation effect on images with very different scenes from the training samples is poor. The experimental results show that U-Net is effective in pedestrian segmentation tasks. However, due to the limitations of training time and hardware capabilities, the segmentation accuracy has room for further improvement. In

addition, the training samples' scene of the dataset is all college students around the university, which may limit the versatility and robustness of the model.

References

[1] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[3] Y. Liu, D. Zhang, and G. Lu, *Region-based image retrieval with high-level semantics*. Citeseer, 2006.

[4] A. Mizushima and R. Lu, "An image segmentation method for apple sorting and grading using support vector machine and otsu's method," *Computers and electronics in agriculture*, vol. 94, pp. 29–37, 2013.

[5] S. Ray and R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," in *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pp. 137–143, Calcutta, India, 1999.

[6] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint arXiv:2001.05566*, 2020.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[11] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.