Impact on Fully Convolutional Network in network structure.

Yang Lu[1,1]
[1] Australian National University.
{u6274652} @anu.edu.au

Abstract. Semantic segmentation is one of the key problem in computer vision. Deep Learning become a popular method in recently years. In a FCN model, convolutional layer is the key part for deep feature extraction. Different network like Vggnet, Alexnet and Resnet architectures do have impact on the model performance. In our work, we implement FCN based on Resnet, the state-of-art network architecture. To show the importance of network architecture, we also compare the model performance between Alexnet, Vggnet, and Resnet.

Keywords: Semantic segmentation, Fully convolutional Network, Resnet

1 Introduction

Image classification, object detection and image semantic segmentation are the three core research issues in computer vision. Semantic segmentation of images is challenging. Image Semantic Segmentation combines traditional tasks of image segmentation and target recognition. It divides the image into a set of semantically meaningful blocks and identifies the class of each segmented block. The result is a pixel-by-pixel segmentation, semantic annotation of images. Currently, semantic segmentation of images is a very active research direction in the field of computer vision and pattern recognition, and it has extensive application value in many fields.

However, the task of semantic segmentation of images is a very challenging problem, the summary of other tasks is similar. The difficulties are mainly reflected in the following aspects:

1) Object level: The same object, due to different lighting, viewing angle, and distance, the images taken will also be very different;

2) Class hierarchy: The difficulties faced at the class level mainly come from two aspects, namely, the dissimilarity between the objects within the class and the similarity between the objects between the classes;

3) Background level: Usually a clean background helps to achieve semantic segmentation of images. However, the background in actual scenes is often intricate, and this complexity also greatly increases the difficulty of image semantic segmentation.

Between 2005 and 2012, PASCAL VOC (Pattern Analysis, Statistical modelling and Computatative Learning, Visual Object Classes) [1] publishes databases for image classification, object detection, or semantic segmentation each year, and runs algorithm competitions on corresponding databases. Since 2007, the category data of the database has been fixed at 20, mainly involving objects common in daily life, including: aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, dining table, dog, horse, Motobike, person, pot ted plant, sheep, sofa, train, and tv/monitor.

The images currently used to measure semantic segmentation algorithms are usually from PASCAL VOC 2012. The original competition database provided 1464 training images with annotation information. In 2014, the literature [2] relabeled nearly 10,000 images and raised the training sample data to 10582. PASCAL VOCs are multi-label databases, each containing one or more objects. The scale of the object varies greatly, the background of the image is complex, and there are often

occlusion phenomena among different objects in the same picture, so the semantic segmentation of the image is more difficult. PASCAL VOC is the most well-known database for evaluating image semantic segmentation algorithms. It poses a great challenge to semantic segmentation algorithms, and greatly promotes the development of semantic segmentation research.

2 Method

The core idea of the Semantic Segmentation Full Convolutional Network (FCN) is to establish a "full convolutional" network, enter arbitrary dimensions, generate an output of corresponding dimensions after efficient reasoning and learning, and learn pixel-to-pixel mapping. Here we implement FCN with 32x unsampled skip architecture[3], see Figure1.

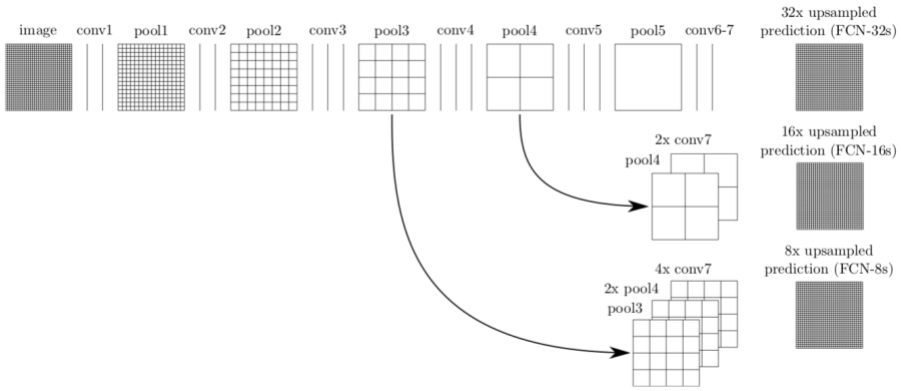

Fig. 1. This figure shows a general FCN architecture[3]. The first section is convolution layer including convolution and max-pooling. The section part is transpose convolution with skip connection.

FCN is the starting point of the semantic segmentation deep learning model, which defines a generic model framework for image semantic segmentation: adapting basic classification networks like Alex Net [4], VGG-Net [5], GoogLeNet [6] extract coarse features, then perform classification prediction to generate segmentation maps, and finally optimize the output of the resulting segmented images. Long and Shelhamer propose Fully Convolutional Neural Network [3] in 2015[but ResNet [7] was published in 2016. Here we evaluate different performance on AlexNet, VGGNet and ResNet.

2.1 AlexNet

In 2012 Krizhevsky won the ILSVRC 2012 image classification competition using a convolutional neural network and proposed the AlexNet model [2] (thesis address). This article, with its many innovative methods, prompted a wave of post-neural network research. The proposed AlexNet network has a landmark significance for convolutional neural networks. Compared with the improvement of LeNet-5[8], there are the following points: Data enhancement, horizontal flip, Random cropping and translation, Color light conversion, Dropout, ReLU, LRN.

It is notable that the Dropout method, like data enhancement, prevents overfitting. Simply put, dropout can drop neurons from the network with a certain probability but here we do not use fully connection. ReLu has some excellent characteristics, which can introduce sparseness while introducing nonlinearity to the network. Sparsity can selectively activate and distribute the activation of neurons. It can learn relatively sparse features and play an automatic dissociation effect. In addition, the derivative of ReLu is larger than 0, the derivative of the function is 1. This feature can guarantee that the gradient does not decay when the input is greater than 0, so as to avoid or suppress the disappearance of the gradient during network training. The convergence speed of the network model will be Relatively stable [10]. Overlapping Pooling: Overlapping means that there is overlap, that is, Pooling's step size is smaller than the corresponding side of Pooling Kernel. This strategy contributes 0.3% Top-5 error rate. Multi-GPU parallelism: This is too important. After entering the pit, it is found

that deep learning is really a discipline of "Alchemy". Thanks to the development of computer hardware, GPU is probably over an order of magnitude faster than CPU in my own training. Can greatly speed up network training.

2.2 VGGNet

VGGNet was proposed by the Visual Geometry Group of Oxford University and is the first place in the ILSVRC-2014 positioning task and second in the classification task. Its outstanding contribution is to prove that using a small convolution (3 * 3), increase the network depth can effectively enhance the model's effect, and VGGNet has good generalization ability for other data sets.

Today, convolutional neural networks have become a common tool in the field of computer vision, so many people have tried to improve the AlexNet proposed in 2012 to achieve better results. For example, the best-performing ZFNet[9] in ILSVRC-2013 uses a smaller reconstructive window size and a smaller stride in the first convolutional layer. Another strategy is to intensively train and test over the entire image in multiple scales. VGGNet emphasizes another important aspect of convolutional neural network design—depth.

In this paper we evaluate VGG net based on Pytorch pretrained VGG16 network. And then we replace the Fully connected layer with a Fully convolution layer.

2.3 Residual Network

The most fundamental motivation for ResNet is the degradation problem. When the depth of the model is deepened, the error rate is improved. However, the depth of the model is deepened and the learning ability is enhanced. Therefore, the deeper model should not produce a shallower model than it is. Higher error rate. The reason for this "degradation" problem is attributed to the optimization problem. When the model becomes more complex, the optimization of SGD becomes more difficult, resulting in the model failing to achieve good learning results. The author proposes a Residual structure in Figure1.

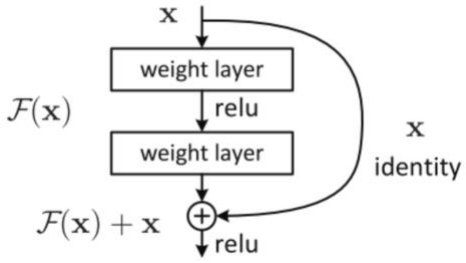

Fig. 2. A normal residual block in Resnet, it shows the signal path.

That is, adding an identity mapping, the original required learning function $H(x)$ is converted to $F(x) + x$, and the author thinks that the two expressions have the same effect, but the optimization difficulty is not the same. The author assumes that the optimization of $F(x)$ will be much simpler than $H(x)$. This idea is also derived from the residual vector coding in image processing. Through a reformulation, a problem is decomposed into multiple scales and direct residual problems, which can be used to optimize the training effect.

The Residual block is implemented by a shortcut connection. The input and output of this block are subjected to an element-wise addition via a shortcut. This simple addition does not add extra parameters and calculations to the network, but it can greatly increase the number of models. Training

speed, improve training effect, and when the number of layers of the model deepens, this simple structure can well solve the problem of degradation.

3 Result and Discussion

Besides the network output in Figure 3, the performance of the semantic segmentation model evaluates in 2 metrics: classification accuracy and mean IoU(Intersection-Over-Union).

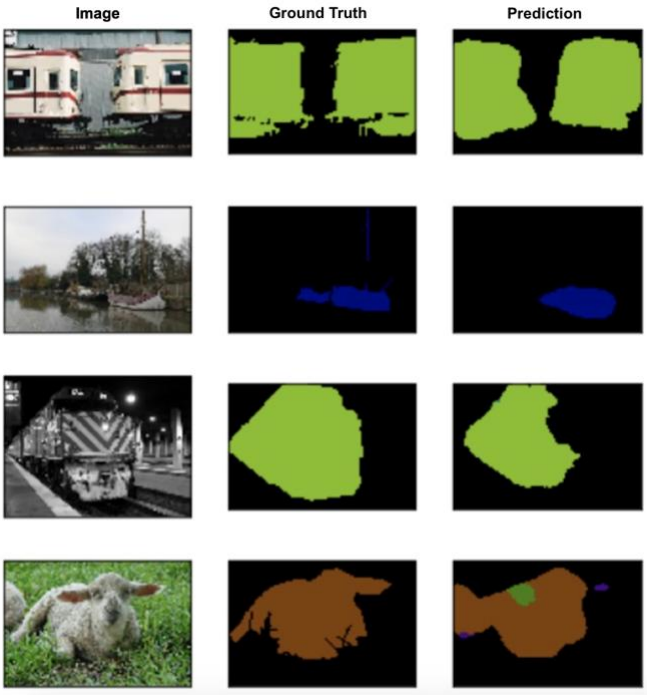

Fig. 3. Shows the network input, ground truth and output. A semantic segmentation problem regard as a classification problem for each pixel.

The purpose of a suitable evaluation matrix for a computer vision problem is to identify the (dis)similarity between the possible solution and the ground-truth presented in a perceptual way. For solving the problem of semantic segmentation, we decide to choose the popular matrix standard Jaccard Index or known as Intersection-Over-Union (IoU) measures, which computed the discrepancy between the ground-truth area and the predicted area, and calculate the average value among all categories.

The IoU score is a common method to evaluate the performance in the semantic segmentation problems. If the images dataset is given, the IoU score will reflect the similarity between the ground-truth area and the predicted area for a specific object presented in images. Figure 4 demonstrate the training process evaluation on NVIDIA Telstra K40 GPU. We also compare the performance on different network architecture based on the result from Long and Shelhamer [3].

Fig. 4. The graph of mean IoU, Accuracy and Training Loss. With the increase of Training IoU and Accuracy, the training loss drop from the beginning toward the end. It finally converges to 0.18.

Table 1. Comparison in different Network Architecture.

	FCN-AlexNet[2]	FCN-VGG16[2]	FCN-Res34[2]
Mean IoU	39.8%	56.0	65.12%
Conv. layers	8	16	34
Parameters	57M	134M	21M
Max Stride	32	32	32

In this table, we compare mean IoU, the number of Convolution layers, Network parameters, and Max from Long and Shelhamer[3]. FCN-AlexNet reach 39.8% accuracy rate. A future work FCN-VGG16 can achieve 56.0%. My Resnet34 implementation with 32x unsample have a 65.12% accuracy rate.

4 Conclusion and Future Work

Semantic segmentation is hard. It may work well in training set but not robust, and lead to ambiguous boundary. Sometimes neural network is efficient but sometimes may over-fitting. Our experiment shows it is meaningful to use a more deeper architecture. In the future work, we hope to use encoder-decoder network structure to get more precise prediction and higher accuracy and compare the result.

References

1. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal- network.org/challenges/VOC/voc2011/workshop/index.html.

2. Pont-Tuset, J., Arbelaez, P., Barron, J. T., Marques, F., & Malik, J. (2017). Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE transactions on pattern analysis and machine intelligence, 39(1), 128-140.

3. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

4. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

5. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

6. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015, June). Going deeper with convolutions. Cvpr.

7. Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in Resnet: generalizing residual architectures. arXiv preprint arXiv:1603.08029.

8. LeCun, Y. (2015). LeNet-5, convolutional neural networks. URL: http://yann. lecun. com/exdb/lenet, 20.

9. Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.