Semantic Segmentation Project

DDA4220/MDS6224/MBI6011 Final Project

Wu Jiacheng 221041013
School of Data Science
Chinese University of Hong Kong, Shenzhen
221041013@cuhk.edu.cn

Chen Fangda 121090029
School of Data Science
Chinese University of Hong Kong, Shenzhen
121090029@cuhk.edu.cn

Sun Jiashu 222051030
School of Medicine
Chinese University of Hong Kong, Shenzhen
222051030@cuhk.edu.cn

Abstract

Semantic segmentation is one of the most popular areas in computer vision. In this project, we trained UNet and DeepLabv3+ for the segmentation of different parts of human body. Our best model achieved 59.23 mIoU and performed well in head and torso segmentation. We also tried Lovasz loss function to improve the smoothness of the segmentation border.

1 Workload

Contributions of each member (if group project): Wu Jiacheng (30%), Chen Fangda (50%), Sun Jiashu (20%). Wu Jiacheng (40%) conducted the experiments of U-Net and Deeplabv3 as baseline models. Chen Fangda provided improvements with pretrained ResNet50 model and Lovasz loss function. Sun Jiashu investigated the related works to help design the experiments and contributed to the report writing.

2 Introduction

For computer vision(CV) problems, there is a technique called semantic segmentation, which involves dividing the given images into different regions, and then make each region correspond to a particular object or part of an object, helping computers understand and interpret the images just like humans. This technique has now become a very famous tool for many applications, like autonomous driving or medical imaging.

Though semantic segmentation has been improved a lot due to the progress of deep learning, some problems still remain, like occlusions, variations in object appearance and handling large amounts of data.

This report is the workflow about a semantic segmentation question. We intend to use this tech to analyze pictures about humans which is provided by professor. .

3 Related Work

For many years, semantic segmentation has been a important area in CV. Early methods[1] for semantic segmentation are region-based or edge-based, which means these methods firstly extract features from regions or edges of images, then utilize these features to assign semantic labels to each pixel. Nowadays, with the progress of deep learning, new methods based on deep learning have become popular, among them, U-Net, Fully Convolutional Networks(FCNs) methods are what we concerned about.

Hence in this task we use FCN+U-Net as our baseline., and we also tried deeplabv3 and Lovaz loss function.

FCN[2]: FCN (Fully Convolutional Network) is neural network architectures used for semantic segmentation tasks. They produce dense pixel-wise outputs that correspond to a semantic segmentation map of the input image.

U-Net[3]: U-Net is a convolutional neural network architecture consisting of an encoding path that gradually reduces the spatial dimensions of the input image and a decoding path that upsamples the feature maps to produce the segmentation map and is used commonly for image segmentation tasks.

DeepLabv3[4]: DeepLabv3 is a convolutional neural network architecture employing atrous convolution and a spatial pyramid pooling module to capture multi-scale contextual information and produce accurate segmentation maps and it's used for semantic image segmentation tasks.

Lovaz loss function[5]:A method for direct optimization of the mean intersection-over-union loss in neural networks, in the context of semantic image segmentation, based on the convex Lovaz extension of sub- modular losses. The loss is shown to perform better with respect to the Jaccard index measure than the traditionally used cross-entropy loss.

In this report, we aim to address some of these gaps in the literature by proposing a novel approach to semantic segmentation for human image recognizing.

4 Method

In this report, out purpose is to do semantic segmentation on images which are about human beings. The dataset is provided by the professor. Dataset is in "Pascal_seg.zip". In specific, we intend to segment 7 semantic parts including head, torso, upper-arms, lower-arms, upper-legs, lower-legs, and background. For this task, we used U-Net and FCN. U-Net is very good at solving image problems especially for image segmentation. Normally, the structure of U-Net looks like a "U", which contains 2 parts, encoder and decoder, and such a structure is good at getting the context from both local or global, which could make U-Net could work with high accuracy. FCNs consist of a series of convolutional layers that gradually reduce the spatial dimensions of the input image while increasing the number of feature channels. The final layer of the network is a convolutional layer that produces an output tensor with the same spatial dimensions as the input image, but with a different number of channels that correspond to the number of semantic classes in the segmentation task.

The initially combination of the baseline was FCN and U-Net, the outcome is bad.

We then used deeplabv3, but the outcome was still bad.

However, with the pretrained ResNet encoder, we found that the outcome is acceptable.

We also tried Lovasz loss function since we thought ehe characteristics of tight and highly asymmetric connections in limb segmentation, but in the end the outcome was not improved too much.

Figure 1: the architecture of U-Net looks like a "U" letter. Inputted data gets downsampling at first, then gets upsampling, and will be outputted finally.

Figure 2: FCN(Fully Convolutional Networks), is a architecture consists of a downsampling path and a upsampling path. It utilizes locally connected layers (convolution, pooling, and

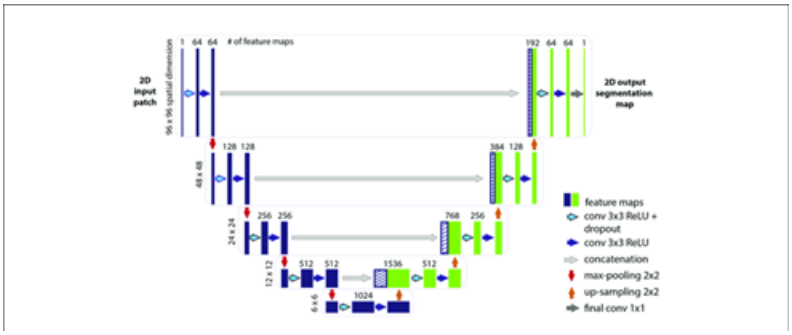

Figure 1: U-Net

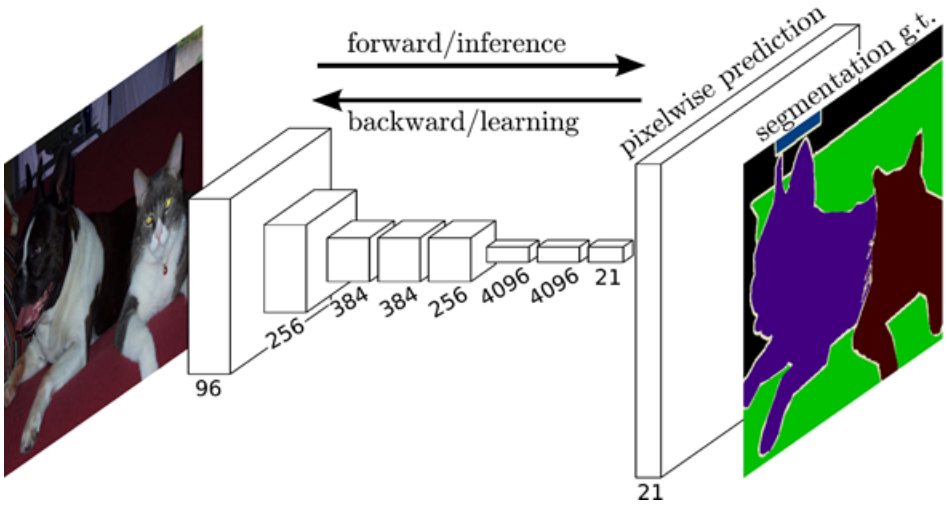

Figure 2: FCN

upsampling) and avoids using dense layers. Such architecture is able to process multiple sizes of images while all connections are local.

Figure 3: DeepLabv3 is a fully Convoltional Neural Network model. This architecture utilizes ResNet models as its backbone along with Atrous Convolution and Atrous Spatial Pyramid Pooling(ASPP) module.

Figure 4: Lovasz loss function (Lovász-Softmax). A method for direct optimization of the mean intersection-over-union loss in neural networks, in the context of semantic image segmentation, based on the convex Lovaz extension of sub- modular losses. The loss is shown to perform better with respect to the Jaccard index measure than the traditionally used cross-entropy loss.

5 Experiments

In summary, we conducted four experiments: 1) the baseline model FCN+U-NET. 2) deeplabV3+. 3) deeplabV3+ with a pretrained ResNet50 encoder. 4) deeplabV3+ with a pretrained ResNet50 encoder and Lovasz loss function.

5.1 Evaluation method

In this semantic segmentation task, we applied two commonly used evaluation metric: Mean Intersection over Union (mIoU) and pixel accuracy.

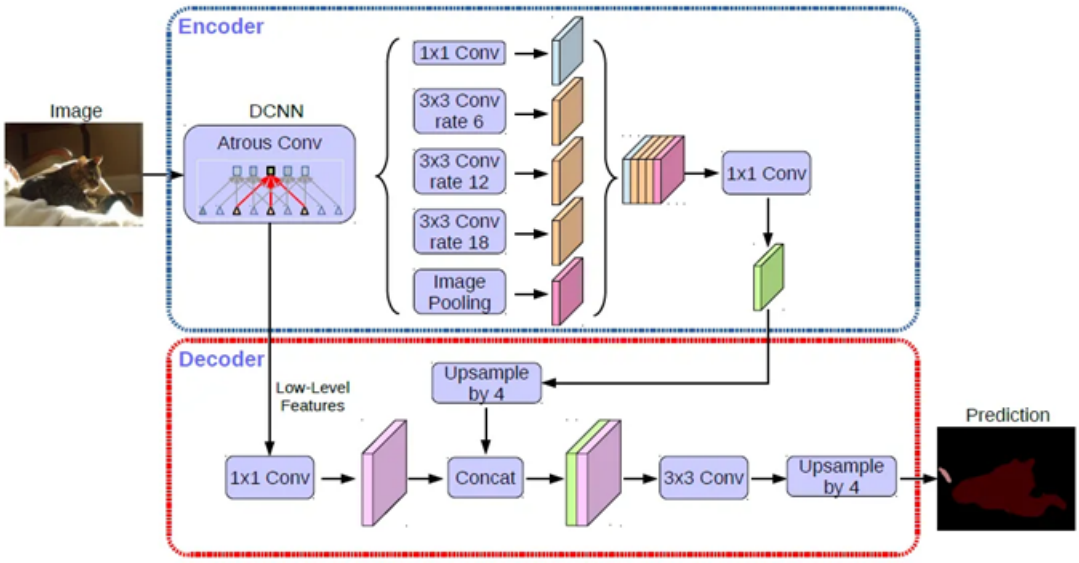

Figure 3: DeepLabv3 + Exends DeepLabv3

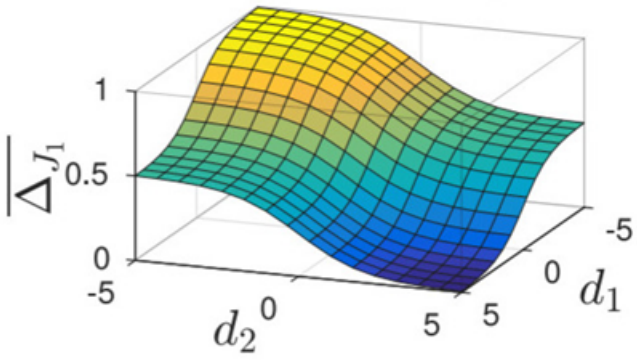

Figure 4: Lovász-Softmax function

Mean Intersection over Union (mIoU): This metric measures the average overlap between the predicted segmentation and the ground truth across all classes. For each class, the Intersection over Union (IoU) is computed as the ratio of the area of intersection between the predicted segmentation and the ground truth to the area of their union. The mIoU is then calculated by averaging the IoU values for all the classes. A higher mIoU indicates better segmentation performance, as it signifies a greater degree of overlap between the predicted and ground truth regions for each class. The mIoU can be calculated as below:

$$IoU(c) = \frac{TP(c)}{TP(c) + FP(c) + FN(c)}$$

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} IoU(c)$$

where:

- $TP(c)$ is the number of true positive pixels for class c.
- $FP(c)$ is the number of false positive pixels for class c.

- $FN(c)$ is the number of false negative pixels for class c.
- C is the total number of classed, here $C = 7$.

Pixel Accuracy: This metric measures the proportion of correctly classified pixels in the segmented image. It is computed as the ratio of the number of correctly classified pixels to the total number of pixels in the image. While pixel accuracy is a simple and intuitive metric, it may not provide a complete picture of the model's performance, especially in cases with a highly imbalanced class distribution. In such scenarios, the model may achieve high pixel accuracy by correctly predicting the dominant class, while performing poorly on the less represented classes. The Pixel Accuracy can be calculated as below:

$$Pixel\ Accuracy = \frac{\sum_{c=1}^{C} TP(c)}{\sum_{c=1}^{C}(TP(c) + FP(c))}$$

5.2 Experimental details

Devices: All the experiments were conducted on a server with an Nvidia Geforce 3090 GPU.

Models: On the way to achieving the highest score on the test set, we conducted four experiments in total.

The first experiment applied a U-Net-structured network as the backbone, which was what we had learned in class. It has an encoding and decoding path, each with 5 convolutional layers with ReLU activations. The encoding path downsamples the input at each stage, reducing its spatial dimensions. Conversely, the decoding path upsamples the feature maps. The number of output channels for the first convolutional layer is set to be 64, and this number doubles after each downsampling operation in the encoder. After the U-Net, two Fully Convolutional (FCN) Heads (decode head and auxiliary head) with dropout rate of 0.1 were added. The whole structure was from the *fcn_unet_s5-d16* configuration provided by MMSegmentation.

The second experiment applied deeplabv3+ model.The backbone was a ResNet50 network and the decoder part used a depthwise separable ASPP head and a FCN auxiliary head. The detailed configuation was stored in *deeplabv3plus_r50-d8*.

We observed severe overfitting during training. So, we consider continuing training on a pre-trained model that has seen a massive amount of data. Specifically, we leveraged an ImageNet pre-trained ResNet50 model as the backbone in deeplabv3+ in the third experiment and end-to-end train it.

The fourth experiment applied Lovasz-Softmax loss based on the third experiment. In the task of limb segmentation, the errors in the border regions of the objects are typically more "difficult" because these regions contain the transitions between different classes. The Lovász-Softmax loss[5], by virtue of the Lovász extension, gives more weight to these difficult errors, may theoretically help to improve the segmentation performance in the border regions.

Data Augmentation: We applied Random Flip and Random Crop as data augmentation. Each image was to be resized so that the size of the short edge was 512. Then, a 512x512 image was randomly cropped from the resized image to be sent to the later pipelines. After that, the image was to flip randomly with a rate of 0.5. Finally, the image would be fed into the network with a batch size of 4.

Training Settings: All the experiments used SGD as the optimizer with $lr = 0.01$, $momentum = 0.9$, and $weight\ decay = 0.0005$. Due to the limited computational resources, each model was trained by 2000 iterations, namely, batches. We believe it was enough since we observed that the loss and the performance did not improve for many iterations.

5.3 Results and analysis

The results for all experiments are shown in table1, 2, 3, 4. The IoU for each class is reported and the mIou is also recorded. The pre-trained model (Table 3, 4) is significantly better (with mIoU 0.7) than those models (Table 1 with mIoU 0.3, 2) trained from scratch because the amount of data in the training set is very limited (1000). Further training on a pre-trained model can utilize the features that are already learned from other datasets, which reduces overfitting since the model does not just

memorize knowledge specific to our training set but learn how to perform well in our training set given the knowledge the pre-trained model has. By observing each class' IoU, the improvement is most substantial for upper arms, upper legs and lower legs, from lower than 0.1 IoU to 40 IoU when a pre-trained model is used, while other classes also see a huge gain in IoU.

We also show the segmentation result for one selected image in the test set (Fig. 5). Visually, the quality of the segmentation coincides with our quantitative result. Models trained from scratch fail to generate a complete result, while models initialized with pre-trained weights produce more complete and correct segmentation, especially in torso, arms and legs.

Class	IoU	Acc
mean	23.09	29.23
background	82.81	97.17
head	45.66	62.2
torso	15.16	17.74
upper-arms	0.21	0.22
lower-arms	17.59	27.04
upper-legs	0.22	0.22
lower-legs	0.0	0.0

Table 1: FCN+U-Net

Class	IoU	Acc
mean	24.31	29.89
background	83.13	96.45
head	39.58	52.61
torso	25.02	34.61
upper-arms	4.35	4.7
lower-arms	10.71	12.52
upper-legs	6.27	7.19
lower-legs	1.12	1.15

Table 2: Deeplabv3+

Class	IoU	Acc
mean	59.32	70.75
background	94.46	97.54
head	82.77	91.1
torso	62.1	78.43
upper-arms	47.98	62.92
lower-arms	46.43	57.9
upper-legs	42.02	56.64
lower-legs	39.51	50.74

Table 3: Deeplabv3+ with pretrained ResNet50

Class	IoU	Acc
mean	58.76	70.38
background	93.81	97.36
head	81.87	88.85
torso	60.49	76.13
upper-arms	47.82	60.42
lower-arms	49.73	63.43
upper-legs	39.75	55.12
lower-legs	37.85	51.38

Table 4: Deeplabv3+ with pretrained ResNet50 and Lovasz loss

Figure 5: From left to right: Original image, FCN+U-Net output, DeepLabV3+ output without pretraining, DeepLabV3+ output with pretraining, and DeepLabV3+ output with both pretraining and Lovász loss.

In addition to the basic setting, we also tried Lovasz loss when training the deeplabv3+. Although the IoU and Acc did not change, we can observe, as shown in figure 5, that the border of different semantic parts got smoother, which was a qualitative improvement.

6 Conclusion

We trained U-Net and DeepLabV3+ for the segmentation of the human body in images. We conducted different experiments based on these two models and different weight initialization methods and recorded the mIoU and pixel accuracy. Training from scratch using the available dataset cannot produce visually desirable results which may be attributed to the small amount of training data and the subsequent overfitting. Models initialized with pre-trained weights performed significantly better visually and quantitatively, which can successfully generate the segmentation. However, our model still does not consider the correlation between different parts of the body. For example, knowing the position of torso may help determine where the leg and arm are in the image. Designing models that can specifically capture such non-local correlation is one of our future works.

References

[1] Yuliia Kniazieva. What is semantic segmentation in computer vision?, 2022.

[2] Thomas Brox Olaf Ronneberger, Philipp Fischer. U-net: Convolutional networks for biomedical image segmentation. 2015.

[3] Trevor Darrell Jonathan Long, Evan Shelhamer. Fully convolutional networks for semantic segmentation, 2014.

[4] Florian Schroff Hartwig Adam Liang-Chieh Chen, George Papandreou. Rethinking atrous convolution for semantic image segmentation, 2017.

[5] Matthew B. Blaschko Maxim Berman, Amal Rannen Triki. The lovaz -softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, 2017.