Semi-supervised Semantic Segmentation with Multi-Task Learning

Luo Li

University College London, WC1E 6BT, UK
zceeidx@ucl.ac.uk

Abstract. Semantic segmentation is a fundamental task in computer vision that involves assigning a label to each pixel in an image. However, obtaining large-scale labeled datasets, especially for semantic segmentation, is a time-consuming and expensive process. Therefore, we proposed a semi-supervised learning method based on the U-Net architecture that leverages unlabelled data for pre-training. We validated the proposed method through experiments, proving the effectiveness of pre-training in learning useful representations of images and thus improving the model's performance in downstream segmentation tasks. The proposed method achieved a mean IoU of around 86.5% on the Oxford-IIIT Pet dataset, beating a baseline model without pre-training by more than 9%. We further examined the utilization of classification labels for multi-task learning. We found that multi-task learning can effectively improve the model's performance in segmentation without pre-training.

1 Introduction

The task of semantic segmentation, which aims to assign semantic labels to every pixel in an image, has long been a fundamental yet widely studied topic in computer vision. Fully Convolutional Neural Networks (FCNs) is one of the traditional methods that achieved great results on several segmentation benchmarks [1-3]. Taking a step further, DeepLab [4] performs spatial pyramid pooling with multi-scale inputs to exploit the contextual information for better segmentation results. The encoder-decoder architecture is also extensively used for semantic segmentation. For example, U-Net [5] is an encoder-decoder architecture based on FCNs which can capture semantic information using the encoder and recover the spatial information using the decoder. The effectiveness of the above models are built on top of sufficient amount of labelled data for supervised training. However, pixel-level labelling of images for semantic segmentation is time-consuming and expensive. This had led to the development of semi-supervised learning techniques that allow models to learn from unlabelled data. For example, [6] uses adversarial learning and unlabeled images to enhance the segmentation model. [7] proposed a few-shot learning like semi-supervised method combining classification and segmentation tasks where the dataset used contains a large amount of images with only classification labels but no segmentation labels.

In this work, we proposed a semi-supervised method based on the U-Net architecture, leveraging unlabelled images for pre-training and labelled images for fine-tuning. We proposed to used masked image modelling for pre-training and multi-task learning with the classification labels of the images to improve the model's performance in semantic segmentation. We aim to examine the effectiveness of pre-training and multi-task learning in helping the model to learn useful representations of the input images under the condition that there exist only a limited amount of labelled data.

2 Methods

2.1 Segmentation Algorithm: U-Net

U-Net [5], a convolutional neural network architecture original designed for biomedical image segmentation, has been widely used for semantic segmentation tasks. The architecture of U-Net consists of a contracting path (encoder) and a symmetric expanding path (decoder) connected with skip connections, forming a U-shape. The encoder consists of a series of convolutional and maximum pooling layers to extract features by gradually reducing the spatial resolution of the input image while increasing the number of channels. The extracted features are then passed to and concatenated with the decoder blocks through the skip connections, which allow the model to preserve fine-grained information from the input image. The decoder consists of a series of convolutional and upsampling layers used for recovering details from the earlier layers and producing an accurate segmentation mask. In this work, we used a ResNet34 pre-trained on the ImageNet dataset as our encoder, which has 21M parameters. The number of stages used in the encoder was set to be 5, where each stage generate features two times smaller in spatial dimensions than previous one. The model takes 4 input channels (3 for RGB images and 1 additional channel for padding masks) and outputs 3 channels for producing a segmentation mask of 3 classes (foreground, background, unknown).

Compared to other popular segmentation models, U-Net has shown to achieve state-of-the-art segmentation performance in various datasets, while requiring fewer training samples and less computational resources. This was first shown in the original paper of U-Net by winning the ISBI challenge 2015 [5] using U-Net. Later in 2018, [8] achieved the highest scores in the online leaderboard of the Medical Segmentation Decathlon challenge using a U-Net based method. Despite its great performance in medical segmentation tasks, [9] demonstrated the power of U-Net in person segmentation using a top view dataset, outperforming the Fully Convolutional Neural Network based methods.

2.2 Semi-supervised Learning

Pre-train: Self-supervised Learning Given a large number of unlabelled data, we proposed to pre-train the model on a self-supervised task, Masked Image Modeling (MIM). We expect MIM to help the model learn more robust representations of image features and spatial relationships between objects, which can improve the model's performance on the downstream segmentation task. Masked Image Modeling is a widely used self-supervised pre-training task in computer vision, which asks the model to predict the masked or perturbed parts of an image based on the surrounding pixels. In this work, we perturbed the input images by randomly selecting 5 small patches of each image and randomly rotating each patch by an angle of 0, 90, 180, or 270 degrees, as shown in Figure 1. Then, we fed the perturbed images into the model and asked it to predict the pixel values of the original images. With RGB images, the model needs to have 3 output channels. Since the downstream segmentation task in this work also requires 3 output channels (predicting trimaps with 3 labels), we consider this pre-training task would be beneficial.

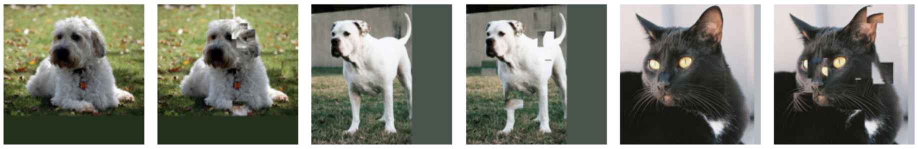

Fig. 1. Three sets of original-perturbed images used for pre-training.

Fine-Tune: Supervised Learning After pre-training the model using unlabelled data, we then fine-tuned the model with labelled data on the downstream task, semantic segmentation. With the pre-trained weights, the model was further trained through supervised learning with the original images as input and segmentation trimaps as output, as shown in Figure 2. The pre-training (self-supervised) and fine-tuning (supervised) parts of the framework therefore combine to form the proposed semi-supervised method for semantic segmentation.

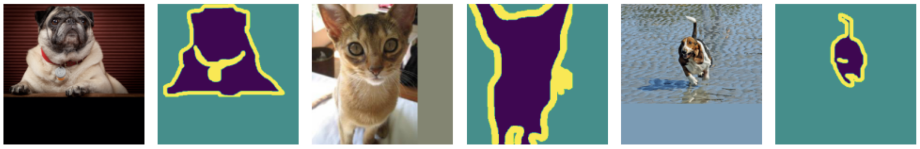

Fig. 2. Three sets of image-segmentation mask used for fine-tuning.

2.3 Multi-task Learning: Classification Head

On top of the proposed semi-supervised method, we further experimented with leveraging the classification labels of the images to perform multi-task learning.

By asking the model to perform both semantic segmentation and binary classification simultaneously, distinguishing between cats and dogs, we expect the model to learn more generalized and robust feature representations of the input images. Since cats and dogs have different characteristics (shape, size, postures, facial features, and etc.), we consider that by learning to classify the input image as cat or dog can effectively improve the model's performance on producing a more accurate segmentation mask, even with a small amount of labelled data.

To perform multi-task learning, we introduced an additional classification head on top of the encoder. The classification head consists of an average pooling layer, a dropout layer, and an auxiliary output layer. During training, we computed the segmentation and classification losses separately and combined them using pre-defined weights (1.0 for segmentation and 0.1 for classification). Then, the weighted sum was used as the total loss for updating the model's parameters through backpropagation.

3 Experiments

3.1 Dataset and Splits

We trained and evaluated the proposed method on the Oxford-IIIT Pet Dataset [4]. It contains 7390 images of 37 distinct categories of cats and dogs, with roughly 200 images for each category. For each image, a segmentation trimap (with three labels: 1-foreground, 2-background, 3-unknown) and two classification labels (one for species: 0-cat, 1-dog and one for the specific category: 0-36).

To perform semi-supervised learning, we need to split the dataset into labelled and unlabelled data. First, we randomly shuffled the the entire dataset. Then, we used the first 40% of the data as labelled data and the rest as unlabelled data. For the labelled data, we further split it into training (80%), validation (5%) and test (15%) sets directly. Such splitting gave us 2956 labelled images (2364 as training, 148 as validation, and 444 images as test), and 4434 unlabelled images. We used the unlabelled data and the training set to perform pre-training and the training set itself for fine-tuning.

3.2 Implementation Details

The images in the dataset have a large variations in scale, therefore, the first step carried out was to perform preprocessing. More specifically, we performed resizing and padding to unify all the images to a target size of 256×256. We first loaded each image using the *Image* module from the *PIL* library. Then, we converted the image to the RGB color space. Next, to preserve the original aspect ratio, we first resized the image to a maximum length (either width or height) of 256 and then padded the image using the edge pixel value to the shape of 256×256. The same preprocessing steps were carried out for the segmentation labels. Finally, we used two arrays to store all the preprocessed images and labels. With all the images and labels having the same shape, we can now load them into mini-batches for training. For pre-training, one more preprocessing step was carried out: image perturbing. As described earlier, each input image was perturbed by randomly rotating 5 small patches of the image.

For loading the data in training, we used a batch size of 64 which is the maximum amount that our 16G GPU can take while giving a fairly good generalization. When training the model during pre-training and fine-tuning, we adopted a technique called learning rate scheduling. During training, the learning rate first increases linearly from 0 to the preset initial learning rate of 0.0003 in the first 800 training steps. This is known as the warmup period which effectively avoids the model from overfitting. Then, the learning rate decreases linearly from 0.0003 back to 0 in the rest of the training steps. This is called the decay period which allows the model to gradually converge to the optimal. The maximum number of training step was set to be 5000 in this work. Furthermore, we also leveraged early stopping to prevent overfitting, where the model stops training when there is no improvement in the validation loss over 5 epochs. Note, a single training step in this work was defined as a single update of the model on a mini-batch of the training data.

During training, the model was evaluated on the validation set after every 50 training steps using the following metrics: mean IoU (Intersection-Over-Union), precision, recall, f1-score, to monitor the training performance. The weights of the model was saved or updated locally if the the mean IoU achieved at the current checkpoint beats the best mean IoU achieved in the previous checkpoints. After training, the model was evaluated over the test set using the same metrics.

3.3 Baseline and Upper Bound Models

Besides the proposed semi-supervised model, we also trained a baseline model and an upper-bound model. Both these two models were trained using labelled data only without self-supervised pre-training. The baseline mode was trained from scratch on the training set only while the upper-bound model was trained on both the unlabelled set (we gave labels to these data) and the training set. In total, the baseline and upper-bound models were trained on 2364 and 6798 labelled images respectively. Similarly, we used the validation set to chose the best models the test set to evaluate the models for comparison.

4 Results

Each model was trained five times with a randomly shuffled training set loader and the averaged metric values were records as shown in Table 1. It was observed that the semi-supervised model outperformed the baseline by a large margin. This validates the effectiveness of the proposed masked image modelling pre-training process on improving the model's performance in the downstream segmentation task. As demonstrated in Table 1, the semi-supervised model was able to achieved similar performance compare to the upper-bound model while using much less labelled data.

Ablation Study An ablation study was also carried out to examine the effectiveness of the proposed multi-task learning method. We trained two sets of models: with pre-training and without pre-training. Within each set, there were two models: one with multi-task learning and one without multi-task learning. As shown in the bottom half of Table 1, we observed that multi-task learning was able to improve the model's performance on segmentation when there was no pre-training. However, the improvement brought by multi-task learning on the pre-trained model was limited.

Table 1. Compare against baseline and upper-bound. (MT: with multi-task learning)

Model	Mean IoU	Precision	Recall	F1-Score
Baseline	0.774 ± 0.006	0.872 ± 0.005	0.851 ± 0.007	0.841 ± 0.004
Semi-Supervised	0.865 ± 0.011	0.921 ± 0.008	0.914 ± 0.005	0.912 ± 0.008
Upper-Bound	0.909 ± 0.004	0.948 ± 0.006	0.948 ± 0.003	0.939 ± 0.005
MT-Baseline	0.802 ± 0.009	0.886 ± 0.004	0.863 ± 0.009	0.856 ± 0.004
MT-Semi-Supervised	0.872 ± 0.014	0.926 ± 0.005	0.921 ± 0.006	0.918 ± 0.009

5 Discussion

As demonstrated above, we validated the effectiveness of our proposed semi-supervised learning method, especially the power of pre-training. Moreover, we found that the proposed multi-task learning method was able to boost the segmentation performance of models that were not pre-trained. The reason that such effect was not observed on pre-trained models is probability due to the the distortion caused by multi-task learning on the representations of images learnt during pre-training. Despite the favourable results obtained in this work, it is strongly encourage to further validate the generalization of the proposed semi-supervised and multi-task learning method on more segmentation datasets. One possible improvement that could be made in future work is to experiment with using dice loss instead of cross-entropy loss in training, where the former was proven to be effective for segmentation tasks.

6 Conclusion

To conclude, we proposed a multi-task semi-supervised method for semantic segmentation based on the U-Net architecture. Through experiments, we demonstrated the power of masked image modeling pre-training on improving the model's performance on downstream segmentation task with a limited amount of labelled data for fine-tuning. The proposed semi-supervised learning model achieved a mean IoU of 86.5% on the Oxford-Pet dataset, outperforming the baseline (without pre-training) by more than 9%. Moreover, we found that multi-task learning effectively improved the model's performance on segmentation by almost 3%, when there is no pre-training.

References

1. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge a retrospective. IJCV (2014)
2. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. (2017)
3. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and stuff classes in context. In: CVPR. (2018)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In: TPAMI (2017)
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention 2015, pp. 234–241 (2015)
6. Hung, W.-C., Tsai, Y.-H, Liou, Y.-T., Lin, Y.-Y., and Yang, M.-H.: Adversarial Learning for Semi-supervised Semantic Segmentation. In: Proceedings of the British Machine Vision Conference (BMVC). (2018)
7. Hong, S., Noh, H., Han, B.: Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation. In: NeurIPS 2015. (2015)
8. Isensee, F. et al.: nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. In: Handels, H., Deserno, T., Maier, A., Maier-Hein, K., Palm, C., Tolxdorff, T. (eds) Bildverarbeitung für die Medizin 2019. Informatik aktuell. Springer Vieweg, Wiesbaden. (2019)
9. Ahmed, I., Ahmad, M., Khan, F.A., and Asif, M.:Comparison of deep-learning-based segmentation models: Using top view person images. In: IEEE Access 2020, 8, 136361–136373. (2020)