MICCAI2022 Singapore

25th International Conference on Medical Image Computing and Computer Assisted Intervention
September 18-22, 2022
Resorts World Convention Centre Singapore

FLARE 2022

Fast and Low-resource semi-supervised Abdominal oRgan sEgmentation challenge

https://flare22.grand-challenge.org/

https://flare22.grand-challenge.org/

NN-UNet with Noisy Student Training

YaoJian Chen
Xiao Fan
Wubing Wan
Qian Yi Ooi

Abstract

The nnU-Net has quickly become a benchmark in the 3D medical image segmentation, which is a self-adapting framework consisting of 2D and 3D U-Nets. In this competition, we combine the nnU-Net with noisy student training, a semi-supervised learning approach that works well on unlabeled data. Meanwhile, we reduce the paramerters of nn-UNet for low resource comsumption.

1 Introduction

The abdominal cavity is the area of the body between the thorax and pelvis. The abdomen includes the stomach, small and large intestines, pancreas, liver, and gallbladder. These organs are maintained loosely together by connective tissues that allow them to expand and collide. Additionally, the abdomen contains the kidneys, the adrenal glands, the oesophagus, the duodenum, and the spleen. Numerous vital blood arteries, including the aorta, inferior vena cava, and numerous of their lesser branches, pass through the abdomen.

Radiologists use computed tomography (CT) to examine the abdominal organ's form and textural abnormalities. These anomalies are critical biomarkers for quantifying organs, planning surgical procedures, and diagnosing disease. Due to the high cost, time-consuming nature, and operator-dependent nature of diagnostic imaging, fully automated abdominal segmentation from CT scans is the most desirable goal. Nonetheless, it remains an open problem because diverse acquisition techniques, contrast agents, contrast enhancement settings, and scanner resolutions all contribute to variable outcomes.

Over the last decade, different methods for automatic, semi-automated, and interactive organ segmentation have garnered substantial attention. It is difficult to determine which strategies are worth pursuing in research and clinical practise. Additionally, the real performance of the best algorithms available today cannot be determined conclusively, nor can the present segmentations generated by automated computational approaches be compared to the ratings of human expert groups. As a result, clinicians are frequently required to manually demarcate regions of interest for a variety of therapeutic applications.

Preprint. Under review.

Due to the availability of massive annotated datasets and affordable parallel computing resources, recent breakthroughs in machine learning have resulted in a dramatic growth in the number of medical picture segmentation techniques. The purpose of this study was to develop semi-supervised learning utilizing nnU-Net (no-new-UNet), Noisy Student Training and Vision Transformer (ViT), by examining the useful information included in unlabeled cases from the training set. The liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, oesophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum were all segmentation targets. Along with the standard Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), our assessment metrics took inference speed and resource consumption (GPU, CPU) into account. Additionally, the area under the GPU memory-time curve and the area under the CPU utilisation-time curve were resource-related measures.

2 Methodology

2.1 U-Net for Biomedical Image Segmentation

Deep convolutional networks have outperformed the state of the art in many visual recognition tasks. While convolutional networks have already existed for a long time, their success was limited due to the size of the available training sets and the size of the considered networks. The typical use of convolutional networks is on classification tasks, where the output to an image is a single class label. However, in many visual tasks, especially in biomedical image processing, the desired output should include localization, i.e., a class label is supposed to be assigned to each pixel. Moreover, thousands of training images are usually beyond reach in biomedical tasks.

Ronneberger et al. [5] built upon a more elegant architecture, the so-called fully convolutional network. They modified and extended this architecture such that it works with very few training images and yields more precise segmentations. In order to localize, high resolution features from the contracting path are combined with the upsampled output. A successive convolution layer can then learn to assemble a more precise output based on this information.

One important modification in the architecture is that in the upsampling part we have also a large number of feature channels, which allow the network to propagate context information to higher resolution layers. As a consequence, the expansive path is more or less symmetric to the contracting path, and yields a u-shaped architecture. The network does not have any fully connected layers and only uses the valid part of each convolution, i.e., the segmentation map only contains the pixels, for which the full context is available in the input image. This strategy allows the seamless segmentation of arbitrarily large images by an overlap-tile strategy. To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image. This tiling strategy is important to apply the network to large images, since otherwise the resolution would be limited by the GPU memory.

Ronneberger et al. [5] then demonstrated the application of the u-net to three different segmentation tasks. The first task is the segmentation of neuronal structures in electron microscopic recordings. The u-net (averaged over 7 rotated versions of the input data) achieved without any further pre- or postprocessing a warping error of 0.0003529 and a rand-error of 0.0382. Ronneberger et al. [5] also applied the u-net to a cell segmentation task in light microscopic images. They achieved good IOU as well. According to the experiment, the u-net architecture achieves very good performance on very different biomedical segmentation applications.

2.2 nnU-net ('no-new-UNet')

nnU-Net ('no-new-UNet') has gradually become a benchmark in medical image segmentation challenge. The original U-Net is a successful encoder-decoder network that aggregates both semantic and spatial information via skip connections. nnU-Net integrates a pool of basic U-Net architectures consisting of a 2D U-Net, a 3D-UNet and a U-Net Cascade, as shown in Figure 1. While the 2D and 3D U-Nets generate segmentations at full resolution, the cascade generates low

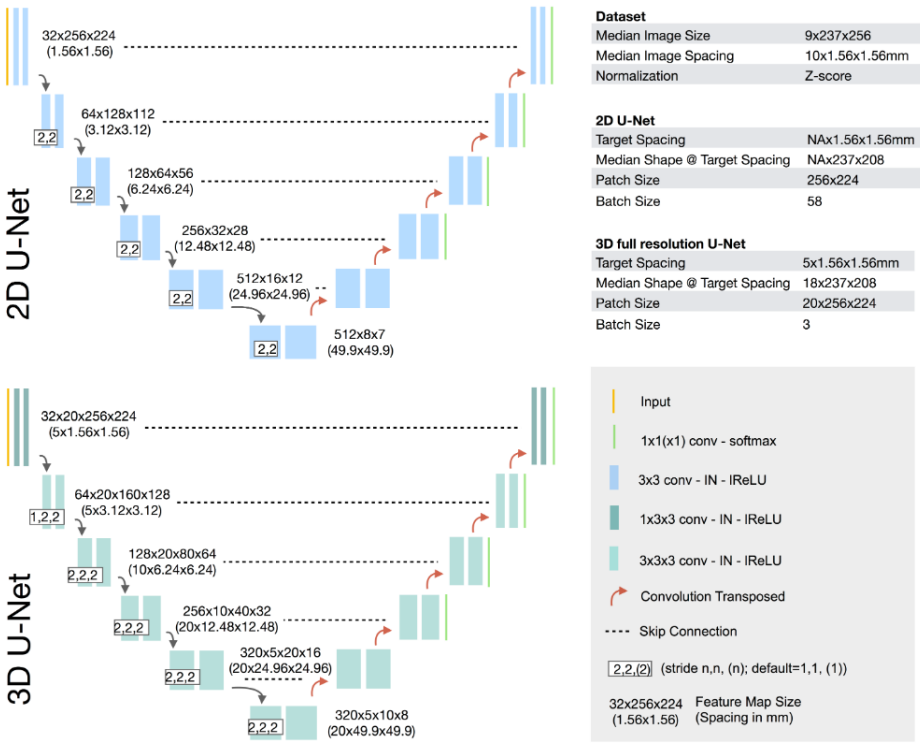

Figure 1: Network Architecture of nnU-Net

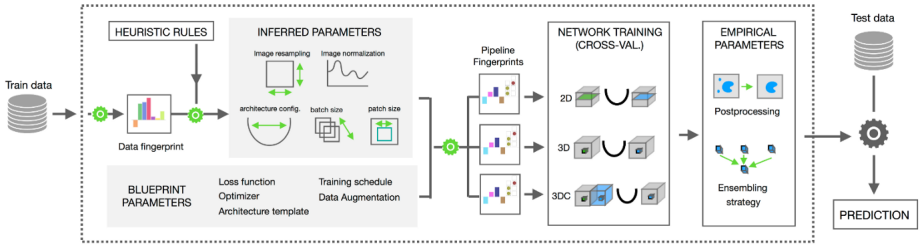

Figure 2: nnU-Net Complete Work Flow

resolution segmentations and subsequently refine them.

2D U-Net The architecture of 2D U-Net is similar to the original U-Net. For each 3D dataset, we crop the data into 2D slices(where we use the plane with the hightest resolution) and train the neural network on these slices.

3D U-Net The 3D U-Net is a common approach for 3D segmentation. However this model suffers from the limitation of GPU memory. When the size of dataset is large, we may only crop the data into small 3D patches and use them as inputs, leading to the loss of contexual information.

U-Net Cascade The U-Net Cascade constitudes two 3D U-Nets. The first 3D U-Net is trained on the down-sampled data and we use the segmentation result(with up-sampling and hot encoding) as the input of the second 3D U-Net.

The nnU-Net pipeline uses heuristic rule to determine the data-dependent hyper-paramters, known as the data fingerprint, to ingest the training data. The blueprint parameters (loss function,

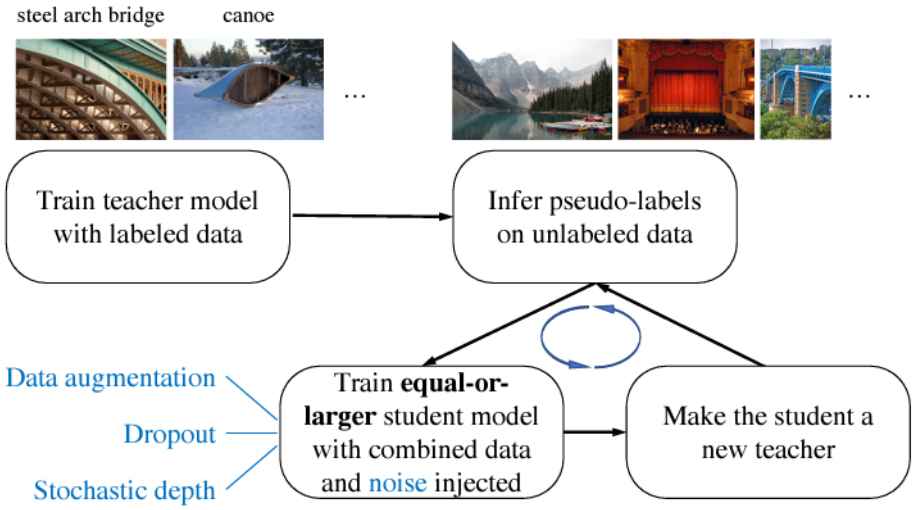

Figure 3: Overview of Noisy Student Training

optimizer,architecture) and inferred parameters (image resampling, normalization, batch and patch size) along with the data fingerprint generate pipeline fingerprints. Pipeline fingerprints produce network training for 2D, 3D and 3D-Cascade U-Net using the hyper-parameters determined so far. The ensemble of different network configuration(s), along with post-processing determines the best average Dice coefficient for the training data. The best configuration will then be used to produce the predictions for the test data. The complete workflow of nnU-Net is as shown in Figure 2.

During inference, all posible combinations of the three models above are ensembled and we choose the ensemble with the highest segmentation score as final model.

2.3 Noisy Student Training

Next, we used Noisy Student Training, a semi-supervised learning approach that works well even when labeled data is abundant. Noisy Student Training achieves 88.4% top1 accuracy on ImageNet, which is 2.0% better than the state-of-the-art model that requires 3.5B weakly labeled Instagram images. On robustness test sets, it improves ImageNet-A top-1 accuracy from 61.0% to 83.7%, reduces ImageNet-C mean corruption error from 45.7 to 28.3, and reduces ImageNet-P mean flip rate from 27.8 to 12.2.

Noisy Student Training improves self-training and distillation in two ways. First, it makes the student larger than, or at least equal to, the teacher so the student can better learn from a larger dataset. Second, it adds noise to the student so the noised student is forced to learn harder from the pseudo labels.

Figure 3 gives an overview of Noisy Student Training. The inputs to the algorithm are both labeled and unlabeled images. We use the labeled images to train a teacher model using the standard cross entropy loss. We then use the teacher model to generate pseudo labels on unlabeled images. The pseudo labels can be soft (a continuous distribution) or hard (a one-hot distribution). We then train a student model which minimizes the combined cross entropy loss on both labeled images and unlabeled images. Finally, we iterate the process by putting back the student as a teacher to generate new pseudo labels and train a new student.

2.4 Preprocessing

We adopt the preprocessing strategy same as nnU-Net.

- Data are cropped to the region of nonzero values.
- Data are resampled to the median voxel spacing of their respective dataset with third order spline interpolation.
- Data are normalized to $[0.5, 99.5]$ percentiles of their intesity values, followed by a z-score normalization.

2.5 Proposed Method

We use nnU-Net for both teacher and student models. The five main steps are:

1. Training a teacher model on the manually labelled data.
2. Generating pseudo labels of the unlabelled data via the teacher model.
3. Training a student model on both manually and pseudo-labelled data.
4. Refine the student model in step 3 on the manually labelled data.
5. Going back to step 2 and replacing the teacher model with the student model for a desired number of iterations.

2.6 Postprocessing

Same as nnU-Net, a connected component analysis is performed on the predicted results. In each predicted class, only the largest connected component for this class is preserved. **Loss Function** We use the summation between dice loss and cross entropy loss because compound loss functions have been proved to be robust in various medical image segmentation task.

A too big model will also cause memory limitation exceedeed, reducing the number of features or levels of convolution will help. Here we choose the 3D network in NN-UNet and set the number of features as 1 to reduce the RAM consumption.

3 Experiment

3.1 Dataset and Evaluation Measures

The FLARE2022 dataset is curated from more than 20 medical groups under the license permission. The training set includes 50 labelled CT scans with pancreas disease and 2000 unlabelled CT scans with liver, kidney, spleen, or pancreas diseases. The validation set includes 50 CT scans with liver, kidney, spleen, or pancreas diseases.

The testing set includes 200 CT scans where 100 cases has liver, kidney, spleen, or pancreas diseases and the other 100 cases has uterine corpus endometrial, urothelial bladder, stomach, sarcomas, or ovarian diseases. All the CT scans only have image information and the center information is not available.

The evaluation measures consist of two accuracy measures: Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), and three running efficiency measures: running time, area under GPU memory-time curve, and area under CPU utilization-time curve. All measures will be used to compute the ranking. Moreover, the GPU memory consumption has a 2 GB tolerance.

3.2 Environment Setting

The development environment and requirement are presented in Table 1.

Table 1: Development Environment and Requirement

Windows/Ubuntu Version	linux 3.10.0-1160.el7.x86 64
CPU	Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz
RAM	376GB
GPU (Number and Type)	Four NVIDIA V100 32G
CUDA Version	10.2
Programming Language	Python 3.7
Deep Learning Framework	Pytorch (Torch 1.10, torchvision 0.2.2)

3.3 Training Protocols

The training protocols are as shown in Table 2.

Table 2: Training Protocols.

Network Initialization	"He" normal initialization
Batch Size	2
Patch Size	40×224×192
Total Epochs	1000
Optimizer	SGD with nesterov momentum ($\mu = 0.99$) and weight decay $3e-05$
Initial Learning Rate (lr)	0.01
Training Time	17 hours

3.4 Implementation Details

We use the 3D-Network in nnU-Net and choose 30 unlabel data in the teacher-student model.

4 Result

Our current result on the leaderbroad are as shown in Figure 4

5 Conclusion

In this

Acknowledgement

The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2022 Challenge has not used any pre-trained models nor additional datasets other than those provided by the organizers. The proposed solution is fully automatic without any manual intervention.

References

[1]Isensee F, Petersen J, Klein A, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation[J]. arXiv preprint arXiv:1809.10486, 2018.

[2] Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10687-10698).

[3] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

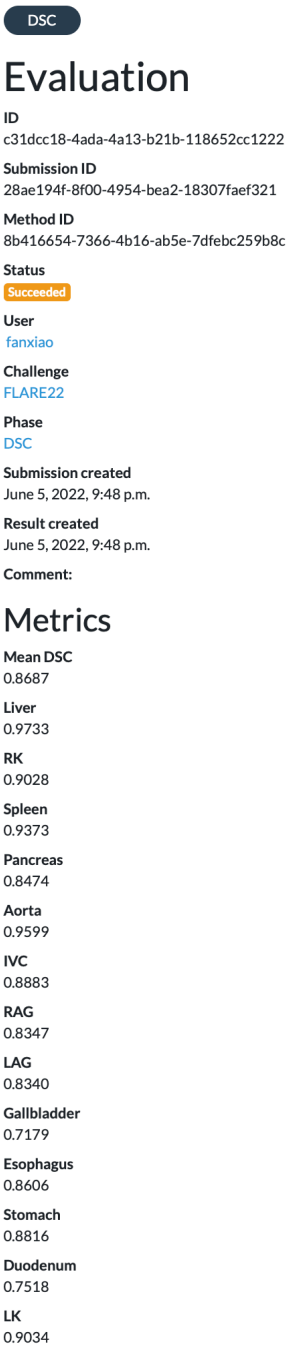

Figure 4: Result on Validation Set (Part I)