POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Multi organ semantic segmentation in CT scans

Laurea Magistrale in Computer Science Engineering - Ingegneria Informatica

Author: Raffaele Spinoni

Advisor: Prof. Daniele Loiacono

Co-advisor: Leonardo Crespi

Academic year: 2021-2022

1. Introduction

Radiation therapy is a medical cancer treatment that uses rays of intense energy to kill the unhealthy tissues.
While planning the radiotherapy is crucial to correctly identify the so-called Organs at Risk (OaRs) to prevent them from being targeted by the radiation therapy. The process of manual delineation of the targets is done by medics, starting from a CT Scan of the patient; for every axial slice the specialist will delineate the various targets pixel by pixel. This laborious effort is not only time-consuming but prone to human errors and results in significant intra- and inter-rater variability [6]. To tackle these limitations, automatic segmentation systems are developed, these systems aim to provide a cheap and scalable solution for treatment planning. The recent developments in the Deep Learning field, and, in particular, in the Image Segmentation task have made it possible to achieve high-quality segmentation results in an end-to-end process.
The OaRs segmentation task is tackled in the context of the Total Marrow and Lymph node Irradiation (TMLI), which is a technique used to irradiate specific parts of the body during treatment. Total Body Irradiation (TBI) was the standard in the past, it didn't need any type of planning because the whole body was hit by the energy beam putting the patients under high doses of radiation and causing late toxicities. TMLI is a novel technique made possible by the latest technologies applied in the medical field, it targets only specific parts of the patients body and, therefore, it needs an accurate planning phase in order to correctly identify the regions to be targeted.
In this context, we designed, developed, and evaluated different segmentation models that are able to automatically segment OaRs from a CT Scan slice. In particular, we compared different settings: we evaluate the efficacy of transfer learning over the standard training "from scratch"; moreover, we evaluated different segmentation settings, considering at first binary segmentation over a single OaRs, then moving to Multiclass segmentation, and, finally, considering the ensemble method Last Layer Feature Fusion to achieve a multiclass segmentation starting from single binary models.

2. Related Works

In recent years, AI has become more integrated with the medical processes, this integration is known under the name of Computer-aided Detection (CAD). While different aspects of the

process are touched by this integration, our contribution is considering the Image Segmentation task, which is the starting point for further planning of the radiation therapy. The *de-facto* standards in the Image Segmentation domain are the Convolutional Neural Networks (CNNs) which are a type of Neural Networks able to learn a spatial hierarchy of features, from low- to high-level patterns.
In the recent literature, CNNs have been widely used for medical image segmentation and lesion detection/classification. For example, in this study [10], the authors developed a dilated convolution network to segment COVID-19 lesions from CT Scan. A dense study has been carried out on tumors and their identification from the medical image: in [5], a segmentation process to identify brain tumors from multimodal MRI images is presented. In the multiclass segmentation settings applied over medical images, we can identify different approaches: experiments have been conducted using 3D volumes, split into batches, and fed to a Neural Network [7]; other researchers analyzed the usage of few contiguous slices as network input [8]. In the literature we can find also adversarial approaches [2], or more complex scenarios, were the segmentation process is split into a Region Of Interest extraction and, after, a binary network segmentation of the identified structure [3].
CNNs are extracting features at different levels and, in general, the deeper the network layer, the more specific the feature extracted. As a consequence, usually, the transfer learning fine-tuning is done over the deeper layers and some of the higher-level layers are frozen. We consider the study done in [9], and we studied the effect of the frozen layers by tweaking the number of them. In the study, the authors analyzed the effect of frozen layer during transfer learning over a target task identical to the source task; they assisted a drop in performance and explained this phenomenon considering that some features may interact with each other in a complex and fragile way, they are *co-adapted*.
Additionally, we considered the multiclass segmentation and evaluate the ensemble method efficacy. Ensemble models combine different models to obtain better results, deep ensemble models combine the advantages of Deep Learning models with the ones of ensemble models. In the literature, different ensemble models have been proposed, based on different fusion approaches. In our work, we consider the Last Layer Feature Fusion ensemble method.

3. Data

The segmentation operations are carried out using Computerized Tomography (CT) Scan images as input data. We consider 2 public datasets referred as *source dataset* where the transferred model are pretrained, and a single private *target dataset* over which we execute the training and evaluation. In particular, the two source datasets are:

- StructSeg2019: a public dataset of CT Scan of 50 patients with annotations of the lung, heart, trachea, esophagus and spinal cord. It consists of 3861 scan/ground truth pairs with size of 512x512.
- SegTHOR2019: another public dataset of CT Scan of 40 patients. It contains ground truth for the heart, aorta, trachea and esophagus. It consists of 7390 total slices with resolution 512x512.

The target dataset is called AUTOMI and contains the CT Scan of 100 patients in DICOM format, each of which contains a variable amount of slices. There are multiple targets labeled (Heart, Esophagus, Liver, Marrow, Spleen, Right Lung, Left Lung, Thyroid, Larynx, Oral Cavity, Brain, ...), but each patient contains only a subset of the total labels. Each couple image/ground truth has the dimension of 512x512. Differently from the source datasets, the ground truth here are binary, meaning that every slice has a ground truth for every target. Another remarkable difference is that the images and masks are rotated of 90 degrees clockwise with respect to the source datasets.

3.1. Preprocessing

The data is preprocessed before being fed to the various networks; the operations executed depend on the type of experiment run. In general, we performed intensity normalization in order to speed up learning and acquire faster convergence; we preprocessed the input slice to obtain values between 0 and 1. While dealing with pretrained models or smaller OaRs we also performed a cropping operation keeping only a central window of dimension 320x320. While

performing fine-tuning we add a rotation operation to account for the difference between the source and target datasets; more specifically, we rotate the couple slice/ground truth of the target dataset by 90 degrees counter-clockwise.

3.2. Data augmentation

Data augmentation is used to increase the amount of data available and increase the generalization ability of the models. We apply different transformations each with 50% probability:

- Elastic Transformation: a non-rigid transformation which, after creating a grid over the image, applies random displacement over the grid intersection points and interpolate the values in between.
- Grid Distortion: another non-rigid transformation that deform objects along the dimensions of the image.
- Rotation: differently from the preprocessing step, we apply a small rotation with an angle chosen randomly between the extremes [-10:10] degrees.

In general, we apply all these augmentations; the exception being the training of *small organs* where only the rotation transformation is applied.

4. Method

The DICOM slices of the patients from the target dataset are ordered over the axial plane from the head to the legs and fed to the various nets.

4.1. Models

The architectures used in our work are Unet, SE-ResUnet and DeepLabV3. Unet and SE-ResUnet share the encoder-decoder structure, while the second uses a special block that weights differently various features (Squeeze and excitement block) [4]. The DeepLabV3, instead, contains the atrous convolution in a pyramid structure, used to avoid too many down-sample and the loss of spatial resolution [1].

4.2. Small Organs

When dealing with hard-to-segment OaRs that occupy a small section of the input slice, we developed a specific segmentation process. The input slice is firstly cropped to a size of 320x320 and fed to a *coarse net* which generates a coarse segmentation, after this step, a smaller window of the input slice is retrieved and stacked together with the last layer context provided by the coarse net. This couple is fed to a *refined net* which outputs a finer segmentation. This process allows for better segmentation performances, thanks to the two segmentations in cascade. The smaller window retrieved both in the input slice and in the context layer is computed by a *window retrieval component* which works in two ways:

- Hard-coded: the central position of the smaller window to be cropped is retrieved from a configuration file.
- Smart-window: the central position of the smaller window to be cropped is computed using the coarse prediction. Using a convolution of constant ones, with kernel dimension equal to the size of the window we can retrieve the window with maximum averaged output.

4.3. Last Layer Feature Fusion

In the context of the ensemble methods, we replicated the Last Layer Feature Fusion method in order to compare the results achieved during the training of multiclass networks over multiple targets. This ensemble method uses multiple binary nets trained on a single OaR, for each net we extract the last layer of features, these are concatenated and fed to a 1x1 convolution layer which weights each features and creates the final multiclass segmentation. In particular, the number of features taken from each Unet or SE-ResUnet is 64, and 256 for every DeepLabV3.

4.4. Training and Evaluation

During training, we used the Dice Loss for the binary segmentation process and the Generalized dice loss for the multiclass cases. We train for 50 epochs with a learning rate scheduler that applied an exponential decay with a rate of 0.6 every epoch. We trained with a batch size of 2 to improve generalization performance and provide training stability, we used a training-validation data split of 80%/20%. The experiments were developed in python 3 using the PyTorch library. The evaluation was done using the Dice Similarity coefficient (DSC) and the Jaccard Index, which are both metrics based on the overlap between predictions and ground truths.

5. Experiments

In general, we divided our experimental works into four categories: we firstly created a baseline of models trained over single OaRs, after, we conducted fine-tuning over different Transfer Learning setups. Lately, we consider the Multiclass cases, and we investigate the ensemble methods improvements.

5.1. Binary Network

We trained and tested binary segmentation models over a set of 23 OaRs, 7 of which were trained using the small organ approach. For each of the experiments, we used the Unet model, comparing its results over the different OaRs in order to have an indicator of the complexity of the segmentation task.

OaR/PTV	DSC	Jaccard Idx
Oral Cavity	85.62	75.86
Ribs	78.42	65.26
Heart	91.83	85.78
Liver	91.80	85.77
Intestine	83.96	73.87
Left Lung	96.39	93.53
Right Lung	96.21	93.39
PTV Abdomen	83.79	73.70
PTV Arms	88.90	81.54
PTV Legs	88.28	81.16
PTV Head	72.41	59.73
PTV Chest	72.53	59.40
PTV Total	79.11	67.99
Rectum	78.40	66.09
Stomach	73.26	61.08
Testicles	75.50	63.60

Table 1: Binary net results for standard OaRs and PTVs.

We present in table 1 the result obtained segmenting OaRs with standard Unet. In the table 2 we show the result achieved using the small organ approach, and we compare the two window retrieval modes, namely, Hard-coded and Smart window. As small window dimensions, we use 140x140 for the thyroid, 120x120 for the right and left parotid and 100x100 for the others.

OaR	DSC	Jaccard Idx
Hard-coded Window		
Esophagus	77.06	63.98
Marrow	82.05	72.36
Left Eye	85.99	76.74
Right Eye	83.10	73.58
Left Parotid	76.48	63.33
Right Parotid	76.01	62.63
Thyroid	64.86	50.21
Smart Window		
Esophagus	77.94	64.99
Marrow	83.54	73.16
Left Eye	84.26	74.83
Right Eye	83.85	74.77
Left Parotid	76.61	63.37
Right Parotid	76.01	62.63
Thyroid	66.19	51.53

Table 2: Binary net results for small organs.

5.2. Transfer Learning

The fine-tuning carried out over the pretrained models (over the source datasets) has been addressed using the same models as the pretrained ones (SE-ResUnet and DeepLabV3). We segmented 5 OaRs using binary nets: Lungs, Esophagus, Marrow and Heart. After testing the base case of models trained from scratch and the complete fine-tuning of the pretrained ones, different fine-tuning scenarios were considered: we checked the effect of the frozen layers, using from 0, up to 3 of them; we simulated data scarcity using 20%, 30%, and 50% of the available patients.
We present the results achieved over the OaR Heart, using the DeepLabV3 net and a center crop size of 320x320. For a better comparison, we also present the results of the training from scratch setting, obtained using the same model.

Setting	DSC	Jaccard Idx
From Scratch	87.91	80.47
From Scratch 20%	84.95	76.74
Fine-tuning	87.91	80.47
Fine-tuning - 1 FL	89.82	83.13
Fine-tuning - 2 FL	89.63	82.82
Fine-tuning - 3 FL	88.73	81.70
Fine-tuning - 50%	88.55	81.47
Fine-tuning - 30%	87.54	80.32
Fine-tuning - 20%	86.13	78.58

Table 3: Transfer learning results in the various settings for the OaR Heart.
The percentages refer to the amount of patients used during training, by default it is 100%, FL stands for Frozen Layer, by default its amount is 0.

5.3. Multiclass

The main objective of the Multiclass setting is to compare the performance of the binary nets with the multiclass ones and create a set of baseline results to be used as a comparison with the ensemble methods. We run multiclass experiments over different sets of targets. We consider as a base case the set of five OaRs considered in the transfer learning scenarios, then we moved over different sets of classes, including also the PTV (areas to be targeted by the radiation therapy). In table 4 we present the results achieved segmenting 6 targets using the DeepLabV3 model and a central crop size of 320x320.

Target	DSC	Jaccard Idx
Left Lung	92.09	86.66
Right Lung	91.77	86.29
Heart	84.80	76.25
Marrow	81.03	68.90
Esophagus	61.72	47.17
PTV Total	80.21	68.23

Table 4: Multiclass net results for 6 targets segmented.

5.4. Ensemble Methods

We train and test the ensemble method Last Layer Feature Fusion, which uses a binary net for each target to segment and takes the last layers of features from each of them, then, merges the single results using a convolutional operation to create the final multiclass segmentation.
In the table 5, we presented the results achieved using our binary net trained from scratch over the single targets, and training the last layer feature fusion module. The input size is 512x512 and the model used for the binary nets are DeepLabV3 for the Heart, Unet for the PTV Total, and SE-ResUnet for the other targets.

Target	DSC	Jaccard Idx
Right Lung	96.03	93.43
Left Lung	96.92	94.33
Heart	88.29	81.19
Esophagus	71.94	58.00
Marrow	84.05	73.62
PTV Total	85.26	75.37

Table 5: Last Layer Feature Fusion net results for 6 targets segmented.

6. Conclusion

Looking at the results, from the binary section, we see that the same network can achieve different results over the various targets. This in general is aligned with the complexity that the medics assigned to each organ. We presented our refinement process for smaller organs, and the results show that the developed pipeline increases the segmentation performances of the binary nets. Interestingly, we noticed that the Smart Window approach doesn't really add any improvement over the base case of a hard-coded window retrieval; this phenomenon can be explained by the fact that the features learned by a CNN are position invariant. In the transfer learning scenarios, we have seen the effectiveness of the knowledge transferred from the same task (from a different dataset). More in detail, the effects of using a smaller percentage of available patients is leading to worse performances,

but this effect has a different order of magnitude: in harder-to-segment OaRs, the reduction of the training set could lead to a drop in performances, while on simpler OaRs the results are almost the same. It seems, therefore, that some patients are more important than others during the training of complex OaRs. Focusing on the variable number of frozen layers, we see, in general, a performance drop for every frozen layer. This effect is not always present, and, considering the work done in [9], we suppose that the feature *co-adaptation* is the responsible for the performance loss when present.
In the multiclass settings, the networks lose precision as the number of targets increases. Looking at the ensemble methods, we see a general improvement over the baseline multiclass results. This trend, however, is not always present and, in some cases of high unbalanced classes, the results achieved by the ensemble models over the smaller targets are lower with respect to the baseline results.
As a general conclusion, generated masks are accurate and able to provide automatic labeling results easing the work of specialists. The results achieved by the fine-tuned networks are an indicator of the efficacy of the knowledge transferred from a publicly available dataset over a private one. In most of the cases, the ensemble methods provide a better solution for the problem of multiclass segmentation thanks to the usage of binary networks pretrained over a single target.

References

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

[2] Xue Dong, Yang Lei, Tonghe Wang, and Matthew Thomas. Automatic multiorgan segmentation in thorax ct images using u-net-gan. 2019.

[3] Raul-Ronald Galea, Laura Diosan, Anca Andreica, Loredana Popa, Simona Manole, and Zoltán Bálint. Region-of-interest-based cardiac image segmentation with deep learning. *applied sciences*, 2021.

[4] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2017.

[5] Yan Hu1, Xiang Liu, Xin Wen, Chen Niu, and Yong Xia. Brain tumor segmentation on multimodal mr imaging using multi-level upsampling in decoder. 2019.

[6] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig1, Aditya Nori, Antonio Criminisi, Daniel Rueckert1, and Ben Glocker. Deepmedic for brain tumor segmentation. 2017.

[7] Pawel M., Herve D., Antonio C., and Nicholas A. 3d convolutional neural networks for tumor segmentation using long-range 2d contex. *Computerized Medical Imaging and Graphics*, 2019.

[8] Holger R. Roth, Le Lu, Ari Seff, Kevin M. Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, and Evrim Turkbey. A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations. *PubMed Central*, 2014.

[9] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? 2014.

[10] Jianxiong Zhang, Xuefeng Ding, Dasha Hu, and Yuming Jiang. Semantic segmentation of covid-19 lesions with a multiscale dilated convolutional network. *Nature*, 2022.