Multiple Organs Segmentation in Abdomen CT Scans Using a Cascade of CNNs

Muhammad Usman Akbar[1,2](✉), Shahab Aslani[1,2], Vitorio Murino[1,3], and Diego Sona[1,4]

[1] Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy
muhammad.akbar@iit.it
[2] Department of Electrical, Electronics and Telecommunication Engineering and Naval Architecture, Università degli Studi di Genova, Genoa, Italy
[3] Department of Computer Science, Università di Verona, Verona, Italy
[4] Neuroinformatics Laboratory, Fondazione Bruno Kessler, Trento, Italy

Abstract. Automatic organ segmentation is a vital prerequisite of many clinical application in radiology. The anatomical variability of organs in the abdomen makes it difficult for many methods to obtain good segmentations for all organs. In this paper, we present a particular ensemble of convolutional neural networks, combining technologies that analyze the images with either a local or a global perspective. In particular, we implemented a cascade of models combining the advantages of using local and global processing. We have evaluated our proposed system on CT scan of 30 subjects in a nested cross-validation framework, showing a significant performance improvement if compared with state-of-the-art methods.

Keywords: Deep learning · Ensemble learning · Convolutional neural networks · Medical imaging · Segmentation · Abdomen organs

1 Introduction

Accurate segmentation of abdominal organs is an important preliminary task in many clinical applications, such as computer aided diagnosis systems, computer assisted surgery systems, radiotherapy systems, etc. Manual segmentation is still a standard practice in radiology that is performed slice-by-slice and organ-by-organ. This makes manual segmentation time consuming and a possible source of errors due to both the variability of human expertise and the inherent subjectivity of the expert. For this reason, there exist many semi-automated segmentation tools, which however still require an interaction with an expert that can introduce biases or unacceptable variability.

To overcome this problem various automated techniques were introduced. Most of the approaches were based either on statistical shape models or on

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11751, pp. 509–516, 2019.
https://doi.org/10.1007/978-3-030-30642-7_46

atlases. Statistical shape models work with an estimation of the distribution of target shapes and have proven to be a successful approach [2,11]. Atlas-based approaches try to segment the images based on registered atlases [4,8,10]. Recently, deep convolutional neural networks (CNNs) have proven to be very effective in many tasks including segmentation, outperforming many state-of-the-art traditional approaches.

In general, all deep learning methods work with two different approaches, either they process the full image (in 3D or slice by slice) [1] or they work with a patch based approach where multiple small patches (in 2D or 3D) are processed separately and results are concatenated to reconstruct the segmentation at the original size [10]. Both approaches enjoy pros and cons. The first approach, thanks to its global processing, is good in locating the organs in the whole space while being less precise on the edges of the segmented areas and on small objects. The second approach instead works on local information, having no perception of the overall objects location while being more able in the segmentation of smaller structures and edges.

Our aim is therefore to create a pipeline of different models combining the two above approaches in order to enjoy the advantages of both frameworks. We propose, therefore, an approach based on the combination of three different CNNs resulting in an improved segmentation where each single approach fails.

Organs segmentation is a difficult task because of the complex anatomical variability of all organs. Due to this variability, machine learning approaches would require datasets with a large number of examples, which is an uncommon condition in medical imaging. For this reason, the most recent CNN based approaches to medical imaging segmentation are limited to single specific organs (usually liver). The proposed method, instead, has been tested on a task requiring the segmentation of 13 different organs, controlling the overfitting through a nested cross-validation

The paper is organized as follows. The proposed system is first explained in Sect. 2, together with a description of the used dataset and the experimental setup. In Sect. 3 results of the experiments will be given and discussed comparing the proposed model with state-of-the-art solutions. Finally, some conclusion will be drawn in Sect. 4.

2 Data and Methods

2.1 Abdomen Organ Segmentation Dataset

We used a publicly available dataset[1] [5] which consist of 30 healthy subjects. The data was hand-labeled with 13 classes corresponding to 13 different abdomen organs with various sizes (Spleen, R. Kidney, L. Kidney, Gallbladder, Esophagus, Liver, Stomach, Aorta, Inferior Vena Cava, Portal Vein and Splenic Vein, Pancreas, R Adrenal Gland, L Adrenal Gland). The data is available in Nifty volumes. We unified the axial spacing to 3 mm. For this purpose, interpolation algorithm was used to interpolate the CT and gold standards to unify

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789.

the axial spacing. Gray value was truncated between -350 and 350 because of the complex boundaries of different organs and size was re-sampled to 256×256 while maintaining the voxel spacing of 3 mm. Extra parts of the image with no organ labels present were cropped for image pre-processing. The average number of slices per subject was 140.

2.2 Proposed System

The proposed system consists of three models incorporated in a simple framework. We connected the three models in such a way that first 2 models, exploiting respectively the global and the local information, produce segmentations that are used together with the input image by a third model. In an ensemble learning perspective, instead of using traditional approaches to combine the outcome of multiple models (e.g. majority vote), we learn how to combine the outcome using a further deep model, which exploits the predictions of the two previous models together with the input data to generate a refined segmentation. This model learns how to use the two previous segmentations according to how much trustable they are on each sub-structure of the whole image. For this reason, since the reliability of segmentations is based on the location of all substructures, the third model must be selected among those processing the full image exploiting the global information. The flow diagram of the proposed architecture can be seen in Fig. 1. In the pipeline the three models are referred as P1, P2 and P3 respectively.

In other words, the first two models, exploiting respectively the global and the local information, are used to generate the auxiliary information which is then used by a third model to generate the final prediction. The models used to generate the auxiliary information, are respectively the Fully Convolutional Network (FCN) [9] and the 3D-UNet patch-based model [7]. The segmentation's from these two models are concatenated together with the original input forming a three-channels image, which is then used as input to a third model FC-DenseNet103 [3] to generate the final segmentation.

FCN Model (P1). We used DLTK implementation [7] with residual block consisting of ReLU activation function followed by 3D convolution layer to extract the features. To handle the stride convolution, we added pooling to the input before the addition in the residual unit. For all convolution layers, Kernel size is 3×3 with stride 1 and padding size 1. In the decoder stage, fully convolutional layers were used to target the output probabilities. Features maps learned at each layer were up sampled to the original size and then fed to the up-score unit where the features from encoder are learned to produce the sparse feature map. The kernel size used for up score unit was 1^3 and finally a soft-max layer was used to produce the segmentation.

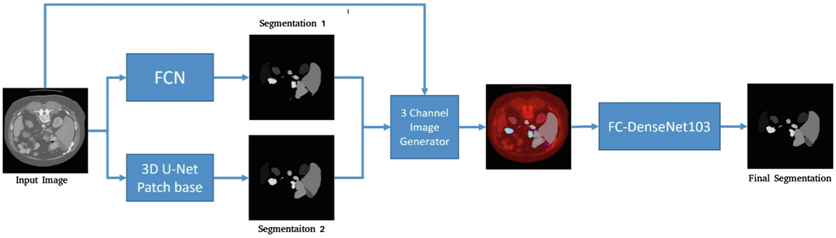

Fig. 1. The entire processing pipeline composed by three CNNs. Input: 256 × 256 gray scale image; Segmentation 1: prediction from FCN model; Segmentation 2: prediction from 3D U-Net Patch Base Model; RGB 3 Channel Image: combination of Segmentation 1, Segmentation 2 and input forming an RGB image; Final Segmentation : prediction from FC-DenseNet103 with the three RGB channels in input.

3D-UNet Patch-Based Model (P2). We used the 3D U-Net patch-based version publicly available[2]. Patch size of 64 is used. Max-pooling operations were performed to reduce the spatial size and high level features were extracted while the bottom block was providing information to the output of the encoder. In the decoder stage two deconvolutional blocks were used to resume the spatial size for the segmented output. In the last stage, convolution and soft-max layers were used to reduce the number of feature maps and to get the probability maps for target objects.

FC-DenseNet103 Model (P3). The input to the last model in the pipeline was a three-channel image composed by P1 and P2 predictions and the original input image. We used the FC-DenseNet103 provided in [3], feature maps are extracted in dense block of transition down layer and use pre-activation layer, where ReLU, convolution, max pooling and Batch normalization were performed on the input slice of 256 × 256. Up-sampling was performed in transition up layer where input was up sampled and concatenated with the skip connections and finally segmentation was calculated using soft-max layer.

2.3 Experimental Setup

FCN model was trained for 20000 iterations with batch size of 8 with tensor-flow. Training was done using Adam optimizer with learning rate of 0.0001. Similarly, 3D-UNet model was trained for 800000 iterations with patch size of 64. Again, Adam Optimizer was used with a learning rate of 0.00001. The last prediction model (FC-DenseNet103) using auxiliary data was trained for 20 epochs with batch size of 8 and tensor-flow as backend. RMSPropOptimizer was used with a learning rate of 0.0001. The proposed system was implemented on NVidia GTX 1080. FCN and FC-DenseNet103 took on average five hours for training while 3d-Unet took on average fifteen hours for training.

[2] https://github.com/zEttOn86/3D-Unet.

For the sake of comparisons we also trained and tested FC-DenseNet103 only using the input images to evaluate the performance improvement due to the auxiliary information. All the mentioned models were evaluated performing a training from scratch, using nested cross-validation with 24 subjects used for training, 2 subjects used for validation and 4 subjects used for testing. For evaluation purpose, we used Dice score to measure the intersection between resulting segmentation and ground truth:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

3 Results and Discussion

The average results (dice scores) for all models are shown in Table 1. The results were determined for all models only using the input image (P1, P2, and P3) and for FC-DenseNet103 also using the auxiliary information (P4). Moreover, the table shows the dice score of four other state-of-the-art methods (IMI, CLS, CNN-sw, FCN), which were top-ranked in the MICCAI challenge providing the dataset [6].

Table 1. Result of varius CNNs. Dice score obtained with Eq. (1) is shown for all models used in our pipeline (P1 is the FCN, P2 is the 3D Unet patch-based, and P3 is the FC-DenseNet103 when considered alone or P4 when considered in the pipeline proposed by the paper. Results of state-of-the-art models (IMI, CLS, CNN-sw and FCN) [6] determined on the same dataset are provided for comparison.

Class	P1	P2	P3	P4	IMI	CLS	CNN-sw	FCN
Spleen	0.856	0.817	0.913	**0.953**	0.919	0.911	0.930	0.936
R. Kidney	0.907	0.902	0.854	**0.934**	0.901	0.893	0.866	0.897
L. Kidney	0.890	0.897	0.813	**0.941**	0.914	0.901	0.911	0.911
Gallbladder	0.543	0.574	0.319	**0.719**	0.604	0.375	0.624	0.613
Esophagus	0.594	0.578	0.624	**0.784**	0.692	0.607	0.662	0.588
Liver	0.920	0.927	0.942	**0.968**	0.948	0.940	0.946	0.949
Stomach	0.757	0.779	0.739	**0.942**	0.805	0.704	0.775	0.764
Aorta	0.840	0.796	0.812	**0.884**	0.857	0.811	0.860	0.870
Inferior Vena Cava	0.782	0.757	0.661	**0.870**	0.828	0.760	0.776	0.758
Portal & Splenic Veins	0.674	0.624	0.498	0.752	**0.754**	0.649	0.567	0.715
Pancreas	0.606	0.613	0.431	**0.832**	0.740	0.643	0.602	0.646
R. Adrenal Gland	0.542	0.513	0.353	**0.752**	0.615	0.557	0.631	0.630
L. Adrenal Gland	0.471	0.462	0.146	**0.702**	0.623	0.582	0.583	0.631

From Table 1, it can be seen that the proposed cascade of CNNs (column P4) performs much better than any other solution, including the state-of-the-art methods top-ranked in the challenge. It clearly indicates the positive effect of auxiliary information provided as further input channels. The effect is reflected by the difference between column P3 and P4 where the same model was used respectively without and with auxiliary information. The results showed significant improvements for some organs, especially small ones like adrenal glands, pancreas, veins, esophagus. This is due to the particular combination of models. Indeed, FCN (P1) and Unet (P2) work at different granularity. The first model works mostly on the global information of the whole image, hence, it better locates the specific organs, while the second model works with local information, being more able to segment on smaller structures and edges. The FC-DenseNet103 model while being a weaker model, thanks to the auxiliary information, it can learn how to use the segmentations provided by the two previous models, refining its own segmentation based on the input images. From an alternative perspective, the FC-DenseNet103 model learns how to cleverly combine the results coming from the ensemble of two other models. In another way, it can be considered an advanced voting approach integrated by the original input as auxiliary information.

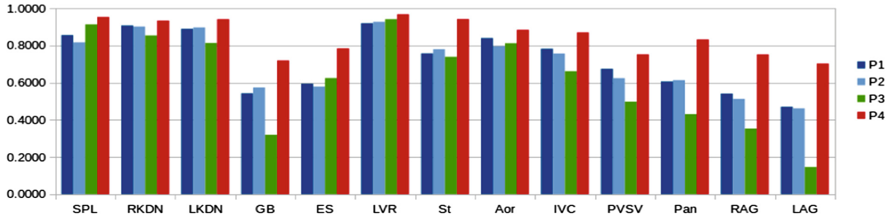

Fig. 2. Comparison of different models average dice used in proposed technique for all class labels.

The bar chart in Fig. 2 is a graphical representation of the results in Table 1 showing the dice score for all elements in the proposed cascade. It can be seen that the entire pipeline (P4), i.e., FC-DenseNet103 with auxiliary information is dominating all classes while the same model (P3) without auxiliary information has sometimes very poor performance. Thus, providing the auxiliary information to the model proved to be beneficial. In order to have a clearer understanding of the processing, the sample results depicted in Fig. 3 show that predictions of P1 and P2 are sometimes affected by small mistakes worsening the performance. However, the third model provided with auxiliary information is able to identify and correct the mistakes producing more accurate predictions. Interestingly, a collateral result is that the processing time of the third model with auxiliary information is reduced as the model is able to learn easily and quickly as compared to when it is not provided with auxiliary information.

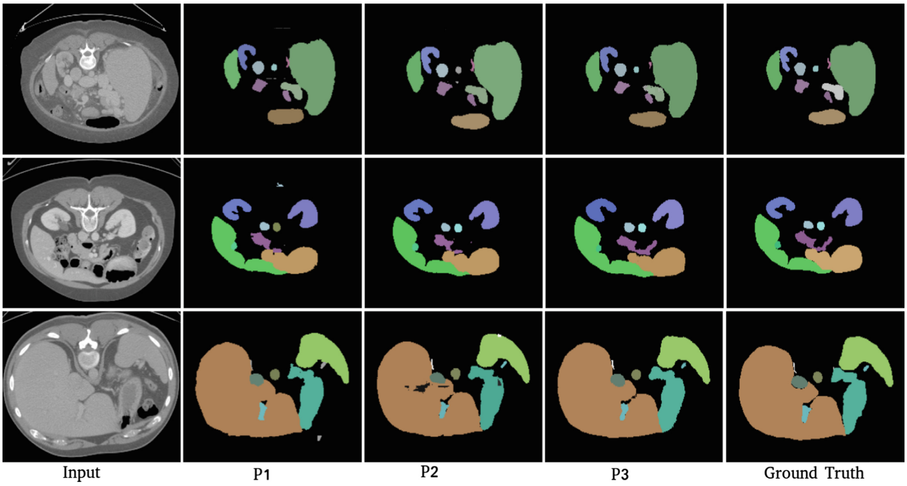

Fig. 3. Examples of Resulting Predictions for three different subjects with Ground Truth

4 Conclusion

In this paper, we proposed an architecture built upon the combination of three different models solving quite efficiently a segmentation task. The three models are connected in such a way that first two models help the third one to produce better segmentation. This is obtained providing the preliminary segmentation from the first two models as auxiliary information to the third one. The relevant component here is the difference in the approaches used by the first two models. The first one segments the organs processing the whole window in one step. This makes the model aware of the location of different organs, losing the precision on small structures and edges. The second method on the contrary is patch-based, hence, it works on local information, making it better when processing small structures and edges.

Giving the outcome of these two models as auxiliary information to a third model helps the system to preserve the positive aspects of all segmentations while ignoring the negative traits. This can be observed from the results in Sect. 3 where it is evident that adding the auxiliary information results in a significant improvement of the segmentation accuracy.

References

1. Aslani, S., et al.: Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage **196**, 1–15 (2019)
2. Heimann, T., Meinzer, H.P.: Statistical shape models for 3D medical image segmentation: a review. Med. Image Anal. **13**(4), 543–563 (2009)

3. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 11–19 (2017)
4. Karasawa, K., et al.: Multi-atlas pancreas segmentation: atlas selection based on vessel structure. Med. Image Anal. **39**, 18–28 (2017)
5. Landman, B., et al.: Multi-modal learning from unpaired images: application to multi-organ segmentation in CT and MRI. In: 2015 MICCAI Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge (2015) https://doi.org/10.7303/syn3193805
6. Larsson, M., Zhang, Y., Kahl, F.: Robust abdominal organ segmentation using regional convolutional neural networks. Appl. Soft Comput. **70**, 465–471 (2018)
7. Pawlowski, N., et al.: DLTK: state of the art reference implementations for deep learning on medical images. In: Medical Imaging Meet NIPS Workshop (2017)
8. Tong, T., et al.: Discriminative dictionary learning for abdominal multi-organ segmentation. Med. Image Anal. **23**(1), 92–104 (2015)
9. Valindria, V.V., et al.: Multi-modal learning from unpaired images: application to multi-organ segmentation in CT and MRI. In: IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2018)
10. Wang, Z., et al.: Geodesic patch-based segmentation. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 666–673. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_83
11. Wimmer, A., Soza, G., Hornegger, J.: A generic probabilistic active shape model for organ segmentation. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 26–33. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04271-3_4